# The Best Subset In Validation Algorithm: Testing Political Scientific Theory Via Predictive Analytics

By

© 2017

Benjamin Rogers

M.A., University of Kansas, 2015

B.S., Illinois State University 2012

Submitted to the graduate degree program in Political Science and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chair: Mark R. Joslyn

Donald P. Haider-Markel

Alesha E. Doan

Gary M. Reich

Clarence E. Lang Jr

Date Defended: 31 May 2017

The dissertation committee for Benjamin Rogers certifies that this is the approved version of the following dissertation:

# The Best Subset In Validation Algorithm: Testing Political Scientific Theory Via Predictive Analytics

Chair: Mark Joslyn

Date Approved: 31 May 2017

Abstract: The difficulties arising from using statistical significance to demonstrate the truth and utility of a hypothesis have been known for some time (Cohen 1994). To develop a more rigorous conception of substantive significance, a new means of determining substantive significance is presented, and a new technique for its implementation is demonstrated. The predictive approach, as differentiated from the explanatory approach (in which significance testing is grounded), concentrates on best determining the value of observations that a given model has not yet seen (Shmueli 2010). The technique for implementation, the best subset in validation algorithm (or BeSIVa), attempts to make the best prediction possible using all available observations. Dividing observations into two separate datasets, training data used for modeling and test data which determines the quality of models at making predictions, BeSiVa tries to best predict predict a dependent variable using a randomly selected test set. BeSiVa is applied to an old question, the choice to vote, as well as two new ones: innumeracy on the proportion of minorities in the United States (Alba, Rumbaut, and Marotz 2005) and support for Donald Trump during his 2016 presidential run. When 656 variables from the GSS were provided to determine if the algorithm regularly selected theoretically relevant independent variables to model turnout, BeSiVa selected a theoretically relevant predictor, voting in the last presidential election, each time. Then, a smaller selection of variables that had been theoretically verified as related to turnout in the 2000 presidential election were then provided. From these variables, BeSiVa clearly favored sociological and psychological theories of turnout over the more recent mobilization theory. Having demonstrated how BeSiVa selected relevant independent variables when analyzing turnout, it was applied to newer questions. Innumeracy's theoretical origins were extended, showing how religious identification and financial satisfaction

predicted an individual's ability to estimate minority proportions. BeSiVa also suggested origins for President Trump's support, grounding it in racial resentment, feelings on President Obama, and concerns about security and immigration. The algorithm's tendency to make theoretically grounded models, even when irrelevant independent variables were provided demonstrates its capability at making useful predictive models with relevant predictors.

## Acknowledgements

It has been a long, difficult, and wonderful five years in graduate school, and the only thing that I need to recount now are the people who helped me along the way. There are so many who deserve my thanks that it's impossible to mention everyone, but I've included as many as I can. I'd like to begin by thanking my advisor, Dr. Mark Joslyn, whose counsel and recommendations gave me exactly what I needed to finish. Because of you, Professor, I named it BeSiVa, and because of you, I successfully made the most important leap of my life.

I also want to thank my parents, Dr. Jean Memken and Dr. Michael Rogers, who first introduced me to the academy and who have supported me tirelessly throughout everything I have done as a student. I also would like to thank my siblings, my sister Sarah and my brother Jonathan, who gave me strength and who were there for me when I needed it.

In addition to my advisor and those closest to me, this dissertation would be incomplete without the invaluable contributions of my committee members. Many thanks are due to Dr. Haider-Markel, who taught the first American politics course I ever took, helped me with the first graduate-level research paper I ever wrote, and with whom I collaborated on the first piece I'll ever publish. I'd also like to thank Dr. Gary Reich, who first taught me how to take notes on the literature, and who taught me how to write under pressure. I'd also like to thank Dr. Clarence Lang, a friend during my time as a candidate, and someone who forced me expand and reconsider how to defend my work. And finally, thanks are due to Dr. Alesha Doan, who taught me how researchers collaborate, and who reminded me of the humanity in the social sciences. Each of you have helped to make this dissertation remarkable and I can only offer my thanks in return.

I'd also like to acknowledge the support of a few other members of my department, especially Dr. Paul Johnson, who helped me to find myself in the discipline, and who taught me the language that underlies the BeSiVa algorithm. You gave me a strong foundation to build upon, Professor, as you cannot spell my research without R. I'd also like to thank Dr. Ron Francisco, who taught me that I belonged here. And last, but far from least, I'd like to acknowledge Dr. Britnee Carter; many people reminded me that there's light at the end of the tunnel in graduate school, but seeing your exit proved it.

Just as much as the faculty and family who helped me along the way, I would like to acknowledge the invaluable support of my friends from outside the academy, and my colleagues who are making their way through it now. Thanks are due to Kim Nixon, my first friend at KU, who helped me figure out how to take care of myself when I was on my own. When it comes to my colleagues, it is impossible to recognize everyone, but I need to thank two irreplaceable office mates. Thanks are more than due to Sara Miller, who taught me to be a better teacher and collaborator, and who gave me what I needed to succeed. I also want to thank Bronson Herrera, whose patience and kindness are inextinguishable. My career in the social sciences was only possible thanks to the work I did with Jerry Moon and John Kupka. Gentlemen, you convinced me that political science needed me just as much as it needs you, with no asterisk required. And I'd like to finish by thanking someone who is more than a colleague, who was with me throughout this dissertation, and whose presence has made me better in every way, Yeonjoo Kim. I thank you all.

# Table of Contents

## Introduction:

"We are drowning in data, and starving for knowledge" Rutherford D. Roger (Quoted in Hastie, Tibshiarani, and Friedman 2009)

The contribution of this dissertation is the BeSiVa algorithm, an approach to both testing the substantive contributions of theory and analyzing data, demonstrated through its application to questions such as the choice to vote. The best subset in validation algorithm (BeSiVa, for short) is discussed, implemented, and its results are analyzed, enabling a rigorous approach to testing the substantive contribution of a variable or collections of variables to predicting the dependent variable. It is this focus on prediction, an approach which is usually set aside in favor of explanation, which is novel to political science.

While the focus on prediction is relatively new to political science, the predictive approach, as implemented in the algorithm, provides a new kind of certainty. The algorithm achieves the goal of not only testing the predictive capability of the variables available, but selecting the ones which best predict the dependent variable. The novel aspects of BeSiVa's approach to variable selection revolve around testing a set of models with different variables according to a more intuitive criterion than statistical significance. This way of selecting models involves separating part of the data from the estimation process for the sole purpose of assessing model quality. This algorithmic separation of data, used to optimize the final model's ability to predict new data, represents a different way of approaching questions in political science, granting a level of predictive accuracy and a different way of expressing substantive significance in analyses.

Difficulties in the way that political scientists conduct research, and difficulties likely to arise due to the increasing availability of larger datasets motivate this dissertation. The implicit approach to figuring out if two variables are related in political science revolves around forming a null hypothesis, based on the alternative hypothesis which the researcher would like to test. Once the alternative hypothesis has been stated, the researcher tests the null hypothesis, commonly with a regression, and bases the analysis' findings on the rejection of the null predicated on whether statistical significance is achieved (Gill 1999). Statistical significance testing has become the standard for determining whether a hypothesis is true, despite a collection of problems that new sources of data are likely to exacerbate.

A major difficulty that null hypothesis significance testing is poorly equipped to deal with is the increased collection and availability of large datasets. Large sources of data are becoming increasingly available in political science, enabling the testing of previously untestable hypotheses. These hypotheses have become increasingly diverse with data's availability, including online censorship in China, the effects of terror attacks on political ideology, and the role of social media in voter turnout (Monroe et al. 2014). While datasets are growing in size, the tools political scientists use to test hypotheses still revolve around statistical significance, an approach where variables are more likely to achieve significance if there is a lot of data. This test becomes less useful as the size of datasets increase, making significance testing uniquely unsuited when testing hypotheses with large data sets.

Despite the promise of large datasets in revealing social scientific relationships, the tools that are used and taught by political scientists remain stagnant and incapable of dealing with the relationships that may appear in these data. As the size of datasets increases, so does the

likelihood of finding statistical significance, regardless of the likelihood of a relationship between the variables. The increased likelihood of achieving significance is instead due to the reliance on testing strategies that unconditionally equate more data with more certainty. Large data sets transform significance testing into a fruitless exercise for determining whether a relationship between variables exists, as 'any difference can be found to be statistically significant given enough data' (Gill 1999, 658). As the amount of data available to researchers grows, so too must the call for approaches that allow researchers to unlock the potential of large datasets. The BeSiVa algorithm, and the predictive approach, constitute an alternative to significance testing, one that provides a new and alternative means of determining relationships between variables.

The increased availability of large datasets is one of the motivations behind this dissertation, but the need for a new approach to such data is only one of the problems that BeSiVa was designed to circumvent. The difficulties with significance testing, both in application and interpretation have been established within the literature, including the illogic in its underlying assumptions (Pollard and Richardson 1987, Cohen 1994, Gill 1999), difficulty in its application and teaching (Haller and Krauss 2002, Reinhart 2015), and a mechanical over-reliance on its conclusions (Schrodt 2014). It is the purpose of this dissertation to demonstrate an alternative approach to confirming or rejecting hypotheses, using BeSiVa to test multiple hypotheses surrounding a theoretically developed area of research, specifically the choice to vote.

Using a predictive algorithmic approach to test relationships between variables represents a different way for political scientists to determine whether a hypothesis or set of hypotheses are

true. It is also a method, however, that can avoid problems common to significance testing, and the approach that underlies the algorithm was developed in multiple literatures outside political science. Separating part of the data from modeling for testing purposes was originally proposed in management studies (Kurtz 1948), and testing how variables relate using a single dataset, one which created those variables, was maligned in the psychology literature, due to its overestimation of how closely two variables related to one another (Cureton 1950). Kurtz's proposed alternative, cross-validation, was differentiated from other approaches to testing relationships by Moiser, who gave the process its name (1951). The use of cross validation as a criterion for model selection was laid out mathematically in statistics, where it was called the predictive sample reuse method (Stone 1974, Geisser 1975). Cross-validation has been applied in political science, and it was first used on a question of textual classification by Mosteller and Tukey (1968, 1977), who attempted to distinguish between the authors of the federalist papers by their most commonly used words. The use of cross-validation and a predictive approach, especially to select models, is hardly new in the literature, but its application in political science requires further consideration.

Despite its heritage and potential utility for testing hypotheses in political science, cross-validation and the predictive approach's uses within political science remain uncommon. Mosteller and Tukey's exhort researchers to cross-validate as a means of testing models (1977), but the use of such an approach is rarely considered for testing model quality, and if used, tends to be for the purpose of testing a single model rather than making comparisons. The BeSiVa algorithm uses a similar approach, using cross-validation to assess model quality, but it expands the idea to focus on varying collections of variables, rather than the tuning parameters of

advanced models (Kuhn and Johnson 2013). While the approach may differ from the explanatory approach common to political science (Gill 1999), the BeSiVa algorithm represents an expansion of the use of cross-validation to determine which of multiple variables are most useful in predicting a given dependent variable.

## Theory and BeSiVa

Given its rarity in the literature, and the fact that BeSiVa automatically selects variables out of a collection provided by the researcher, a reasonable question remains whether such an approach truly can interact with theory. There is a risk using the data to generate hypotheses, of data-mining, especially if the selection criterion concerns goodness of fit or p-values of the data, something which is condemned (Bartels 1997). This approach, known as stepwise selection, and similar approaches lead to incorrect p-values, biased $R^2$ values and severe collinearity issues (Harrell 1996). The use of cross-validation as a criterion, rather than p-values and $R^2$ in the data that was used to create the model, represents an alternative means of testing model and variable quality, one that sidesteps many of the difficulties associated with stepwise selection.

While stepwise selection methods and data mining are critiqued by statisticians and political scientists alike (Harell 1996, Bartels 1997), the use of cross-validation represents a way of avoiding some of these approaches' drawbacks while obtaining their benefits. Attaining variable selection and a measure of prediction, however, comes with additional complexity, as cross-validation has some differences in its interpretation and application compared to significance testing. It adds a layer of complexity which must be addressed, new terminology and a series of steps that are contradictory to the inferential statistical approach, potentially making a determination of appropriate variables independent of any preordained theory. Despite these

complexities, BeSiVa's use as a means of selecting variables can not abolish the role of theory in research, and the algorithm has qualities that should assist theorists in sifting through explanations.

Even with the algorithm's ability to select variables, its use in social science cannot be accomplished without the work of theorists. Part of why cross-validating as a means of variable selection should be attractive to theorists involves its ability to distinguish between a broad collection of theoretical explanations requiring these explanations to make sense of the algorithm's results. A cross-validation approach can use theoretically relevant predictors to weigh competing causal explanations, and in doing so demonstrates that due to the predictive capability of the variables they favor, some explanations are more capable of predicting future observations than others. Unlike ordinary least squares or logistic regression, which accept whatever variables are provided, cross-validation as it is used in the algorithm selects the variables that make the best prediction. Despite the algorithm's variable selection and predictive capabilities, using an algorithm can not replace theory and enables a more comparative and inductive use of theory through the focus on prediction.

The BeSiVa algorithm may have the ability to select and compare between theoretical explanations, but the algorithmic approach should also be attractive to political scientists due to its accidental parsimony. Given the role of cross-validation choice at the heart of the algorithm, there is no reason that using such an approach should lead to a smaller selection of variables to make better predictions (Browne 2000), especially since the algorithm is purely looking for which variables increase predictive accuracy. The most frequent findings, however, are that 2-4 variables are all that is necessary to gain the highest increase in predictive power, with additional

variables decreasing overall accuracy. Part of this may be due to the fact that the algorithm stops when two variables lead to the same increase in the predictive capability of the model, but this accidental parsimony lends credence to the idea that BeSiVa can compare and contrast the utility of theoretical explanations.

BeSiVa may allow for a competition between theoretical explanations, suggesting that a small number of variables may lead to better predictions, but the main motivation for this dissertation is to expand the use of a test which has rarely been used in political science for this purpose. If cross-validation or algorithmic approaches are considered within the literature, they are frequently only mentioned (Lewis-Beck 2005), or used in a limited manner. The most common use of cross-validation in the literature is to determine the role of tuning parameters, options in advanced models that have no mathematically optimal value (Kuhn and Johnson 2013). Alternatively, they have been used as a means of testing classifiers, especially for textual data (Mosteller and Tukey 1968, Mosteller and Tukey 1977, Yu, Kaufmann, and Diermeier 2008). These approaches to cross-validation are appropriate, but they ignore the possibility of the approach as a predictive tool, and how it can be used in a wider array of situations. This view of validation limits it to the point where the American Journal of Political Science has used the term as synonymous for a replication study (Shields and Goidel 1997, Reed 1997). This treatment of a tool that can be used to create a broad array of predictions, to compare theoretical approaches, and to serve as a means of instantaneous replication (Hindman 2015), represents the overall motivation for this dissertation, with the goal of demonstrating the utility of the BeSiVa algorithm for testing variable and model quality.

The major problem of the literature that this dissertation seeks to alleviate is not

theoretically driven. Rather, it attempts to demonstrate the utility of an algorithm which selects models on a criterion that differs from statistical significance: the model's ability to predict a set of data not used in estimation. The standard approach to testing theory, null hypothesis significance testing, has limitations which the algorithm may ease. The predictive approach also allows a researcher to pare down an untenably large list of variables, even in large datasets, and it enables more concrete explanations of how well a model predicts a dependent variable. These are all benefits of an algorithmic approach to data analysis, and the goal of this dissertation is to demonstrate the implementation of the predictive approach, an approach that needs to be a methodological component of political scientific research. In doing so, it strengthens the discipline's ability to make predictions, adding a layer of certainty and substantive significance to analyses, where before there was only uncertainty.

## Organization of the dissertation

In chapter 1, the BeSiVa algorithm is introduced, explaining how it operates and demonstrating its functionality. This chapter concentrates on describing the rules that explain how the algorithm chooses chooses variables and how that leads to the maximization of prediction, and how well the algorithm is capable of choosing apropos variables from the GSS survey. While elements of the chapter include some elements that were previously considered in Rogers (2014), the algorithm's more counterintuitive elements demand a more in-depth treatment. The chapter also includes a description of the motivations behind the rules that make up the algorithm. This introduction to the algorithm requires a consideration of these elements due to their prior disuse in political science, going against the grain of null hypothesis significance testing to determine whether two variables relate.

The technical details of the algorithm include a new way of looking at how variables relate via a predictive criterion, beginning with the algorithm's first step, separating a selection of observations from creating the model for the purpose of testing how well the model predicts those observations. This selection, known as the test set, is used to determine the quality of the models created, using a criterion called the percent correctly predicted. Separating out a selection of the data is necessary to prevent overfitting, as a fit on the same data used to estimate the models allows quirks in the data to dominate overall trends, leading to questionable conclusions (Cureton 1950, Clark 2004). This focus on the algorithm's origins, as well as how it differs from a typical analysis allows for a discussion of how the algorithm functions, in a manner that reveals how a set of variables predicts a given dependent variable.

Having discussed how the algorithm works, its motivations, functionality and application are established in chapter 2, in comparison to regression analysis and how it is used in political science. This is conducted using multiple approaches to falsification, attempting to determine the consistency and utility of the algorithm's results and comparing them to what significance testing would suggest. The results of the algorithm's findings are compared on the choice to vote in the American context against classical models from the literature concerning the likelihood of voter turnout.

In order to determine the algorithm's ability to determine both the choice of relevant and predictive variables, an area where the theory has been developed was chosen. Voter turnout was a natural area to test BeSiVa, not only demonstrating the algorithm's utility in an area where data is rich, but also due to the development of multiple competing theories of the choice to vote. Turnout is a research area with a long and storied history, and the literature's conclusions on

what makes an individual more likely to turn out to vote are well established (Gerber, Green, and Shachar 2003). For this reason, the algorithm's performance can be falsified in a variety of ways. If BeSiVa chooses theoretically relevant variables, an algorithmic approach can be used in a way that interacts with theoretical explanations, suggesting how useful each explanation is in predicting voter turnout. It is also compared against more conventional theoretical models for the purpose of evaluating how capable these models are at making predictions.

In chapter 3, the algorithm is tested on twoproblems with little precedent in the political scientific literature. Having demonstrated its capability at validating theory and selecting appropriate predictors in chapter 2, the algorithm is used as a means of building theoretical relationships through an inductive process. This approach to determining relationships allows the algorithm to act as a means of discovery for questions where theoretical bases have only begun to be established.

# Chapter 1 Sorting Ideas: An Introduction to the BeSiVa Algorithm, how it works, and what it does

"We take seriously indications or indicators not taken further down the inference trail; we take them seriously but with a pinch of salt labelled "indication only". (Mosteller and Tukey 1977, 40)

## Motivations

This chapter concerns the technical aspects of the BeSiVa algorithm, which is short for

Best Subset in Validation algorithm. The technical aspects include how the algorithm works, the results that arise from its operation, and a discussion of counterintuitive elements of algorithmic and predictive approaches to data analysis. Dividing data into subsets for different purposes, creating massive collections of models, and instantly evaluating these models veers far from the typical approach of political scientists, especially in comparison to the explanatory, theoretically driven approach to testing hypotheses, null hypothesis significance testing or NHST for short. These are explored through a discussion of the purposes behind the algorithm, how the algorithm works, and what happens when it is attempted with data from the General Social Survey. This explanation should demonstrate the utility of such an approach to questions in political science, concentrating on the choice to vote from an algorithmic perspective.

The BeSiVa Algorithm was developed as a way to solve a problem with a dataset provided by a political campaign, with the goal of helping the campaign identify supporters. The campaign, however, had provided an untenable amount of data, with hundreds of variables to consider. There were no guidelines on which of the provided variables should be included, and while some appeared relevant, others, such as whether a voter subscribed to hunting magazines or owned a boat, had questionable relevance. Such data would be difficult to use in more traditional social scientific research, but the potential to predict whether such data could model individuals' support for a candidate was part of the challenge. Another aspect of the difficulty of determining who was a supporter was exacerbated by a second demand in the project, that the model be verified using data that was kept separate from modeling.

The data provided presented a variety of challenges, including hundreds of columns, and the campaign added another wrinkle in their desire for extra model verification. In addition to the

lack of clarity in the data there was a desire to demonstrate the effectiveness of any suggested model on a small subset of the data, which was kept separate from the model estimation. Concerned with how to create a good model according to these criteria, especially the suitability of a particular model, the BeSiVa algorithm was written as a means of satisfying the requirements. Concentrating on finding a way to sort any collection of potential independent variables, the algorithm creates the best prediction of a separate subset of data, which is not included for modeling purposes. This required the creation of a way to state how well the held out data were predicted. The algorithm does this by reviewing the data and looking for variables that do a good job of predicting the dependent variable according to its criterion. The purpose of this chapter is to demonstrate this algorithm, and show why a predictive approach allows for a further exploration of questions that were previously thought to be settled, demonstrating its utility on a classic problem: the choice to vote.

### What Researchers Gain from a Predictive Approach:

Before explaining how the algorithm works, it is necessary to consider why prediction can be useful within the social sciences. The social sciences' response to the question of prediction is usually to ignore it, to arguing against its necessity, or to conflate it with explanation (Shmuelli 2010). In his discussion of the future of political science Shapiro points out that prediction is especially difficult in a variety of situations, and its usage in selecting problems is overrated (2002). Shapiro argues for the theoretically driven approach that makes up the backbone of political science research (King, Keohane, and Verba 1994). The difficulty, however, with condemning prediction with regard to selecting problems involves the fact that theoretically driven approaches are fundamentally prediction-based, and that Shapiro sets up a

dichotomy between prediction and theory that does not exist.

While Shapiro suggests that prediction is overrated for selecting problems, he seems to use it to ignore the difficulties with the way political science methodology is regularly conducted. To Shapiro's credit, he acknowledges the problems with purely theoretical approaches, describing how “too often the difficulty is that the theory is articulated in such capacious manner that some version of it is consistent with every conceivable outcome” (2002, 606). Shapiro's article, however, betrays the discipline's overall comfort with theoretical capaciousness. Even as he condemns theory that predicts conflicting outcomes, the theories that Shapiro cites each make a separate, potentially conflicting prediction about the origins of democratic governance. Shapiro discusses the role of theory and empirics within the context of problem finding, but the conclusion that he comes to suggests that prediction as a whole is beyond the reach of political science. Shapiro's overall conclusion, that difficulty precludes prediction as a guiding force in problem selection or solving, avoiding the use of prediction as a guide to deciding which problems need consideration, is betrayed by his own argument, which implicitly acknowledges the necessity of prediction.

Although Shapiro suggests that prediction is not suitable for discerning which problems should be solved, the examples that he describes implicitly make predictions, suitable for testing using a variety of approaches. Each theoretical argument Shapiro mentions already makes a prediction about the origins of democratic governance, but his argument suggests that they should not be considered as such or compared to one another. These theories each make a prediction, ones which may be tested using an approach that explicitly calls for a comparison of predictions. While explanatory statistics as it is practiced may appear to compare hypotheses, an

explicitly prediction driven approach differs significance testing (Shmuelli 2010), and would allow for a comparison of the predictions made by the theories Shapiro discusses.

Although explanatory statistics allow for a consideration of different hypotheses that arise from theory, a predictive approach provides a separate perspective on theoretical utility, while furthering our understanding of a problem on its own. Indeed, parsing between the different theoretical explanations leads to a set of different predictions about the conditions that give rise to democracy. While Shapiro is correct in that theoretically driven approaches, even in the face of capricious theory are necessary, as without the theory, empirics are blind (2002), empirics may provide advice to theorists working from an inductive perspective (Yom 2015). Such an approach allows for a consideration of research methods that acknowledges the predictions that theories make, explicitly considers the relevance of different, competing theories, and chooses between theoretical approaches to gain the most relevant answer to a given question in a given context.

While Shapiro argues against the use of 'prediction-driven' methods of determining what problems are worth solving, theoretical approaches are fundamentally judged by their ability to make correct predictions. This was the argument raised by Meehl, a psychologist and critic of clinical psychological practice. Meehl discussed how clinicians rely primarily on their own judgment, avoiding the use of actuarial tables, scaled approaches, and more statistically rigorous forms of testing in diagnoses and prescriptions (1954). To Meehl, the contemporary clinical approach discards a large selection of useful tests for determining the health of patients, individual likelihood of success in universities, and recidivism among prisoners, as examples. To Meehl, ignoring the predictive abilities of statistics represented the greatest hindrance to

psychological practice, making psychology less capable of achieving its aims.

In his critique of contemporary psychology, Meehl argued that clinicians and advisory groups' approaches to determining patients' conditions tended to lead to diagnoses that were highly subjective and did not help the patient. This was due to the tendency towards prioritizing individual clinicians' judgements for diagnosing and treating mental illness, shying away from more rigorous and systemic methods of judgment. Similarly, practitioners in political science tend towards approaches that favor theoretical arguments, creating hypotheses built from those arguments, and presenting explanatory statistics that validate those arguments (Achen 2002). Such an approach has merits, but it leaves out alternative statistical approaches such as the predictive approach, a differentiation which Meehl identified and recommended.

By advocating for a more systemic psychological practice, Meehl suggested that clinicians were less capable of diagnosing patients than statistically oriented approaches. While making this suggestion, Meehl also distinguished different ways to use statistical methods. Meehl first discussed the structural or analytic use of statistics, which mirrors the explanatory approach preferred by the social sciences (King, Keohane, and Verba 1994, Shmuelli 2010). In this approach, theoretical presumptions are tested via statistical approaches, but Meehl suggested that it was only one way that statistics could be applied to problems in psychology. In his assessment of explanatory statistics, Meehl did not believe that the structural or analytic statistical approach would be most useful in his discipline's practices.

In his discussion, Meehl suggested an approach that focused partially on the analytic use of statistics, akin to the explanatory usage common in social scientific research. Meehl contrasted this approach, however, with the discriminative or validating use of statistics, which

did not concentrate on testing theoretical presumptions. In the validating approach, which mirrors the predictive approach that Shmuelli would later describe (2010), the main concern is whether a test or approach successfully predicted the outcomes that it was meant to predict (Meehl 1960). With this differentiation, Meehl pointed out that the different approaches could both be useful in psychology, but that a consideration of prediction separate from explanation was necessary to make further advances.

By differentiating between analytic and validating statistical approaches, Meehl illustrated the limitations and necessities of altering how statistics were applied in the psychological perspective. The utility of his approach arises from the possibility that statistics may be useful in creating new diagnostic tools for easing mental illness and determining the fitness of individuals in institutional contexts. But Meehl's critiques eschew the necessity of any human touch. Clinician, after all, may detect when a subject is lying, or worse, optimizing their responses to obtain drugs, evade detection, or wrongly gain access to an institution, making a predictive or actuarial approach problematic. In this sense, although a more actuarial approach might provide a standardized response, the clinician remains necessary for diagnosis, administration, adjustment. While this does not invalidate Meehl's argument, it does suggest the continued necessity of the individuals and practices he argues against, an indicator of the need for a balanced approach in diagnosis and assessment.

Despite the convincing need for a more systemic approach in psychological practice, the role of clinicians cannot be denied, especially given the rigidity of actuarial testing. Although Meehl's work suggests that the correct response is to adjust and reconsider the tests, the clinician remains necessary due to the chaos intrinsic to studying and helping individuals. Similarly, those

who focus solely on prediction, who hold it up as the sole arbiter of truth leave out something very important. And yet, even with Meehl and Grove's overzealousness in condemning clinical practice, eschewing the predictive approach in favor of explanation and subjectivity leaves a social science, psychology, for the worse. Similarly, political science needs its theorists, but it also needs explicitly predictive approaches as a means of keeping the theorists honest, and to help their practice whenever possible.

While it is difficult to imagine a discipline like psychology without its clinicians, or political science without theorists, the primacy explanatory statistics in validating theory leaves many questions open. By comparison, a methodology that incorporates prediction into its findings can help fill in parts of a puzzle that the explanatory, inferential statistical approach fails to answer. Prediction is akin to explanatory approaches by allowing for the determination of relevance, but its relevance may be considered a form of substantive significance, something explanatory statistics tends to avoid. Returning to Shapiro's example, if a variable isn't substantively significant, it may help to explain the origins of democracy, but when attempting to predict the rise of democratic governments, the variable is not needed. This substantive significance represents a differentiation between the findings of preditictive statistics and explanatory statistics, and it is necessary to consider both to fully substantiate a finding and its importance. For example, when a social scientist attempts to determine the relationship between two concepts, the common approach is to determine the statistical significance of the relationship in an attempt to determine if x explains y, assuming that there is a relationship if there is statistical significance (Lewis-Beck 2005). This approach ignores the fact that statistical significance should be joined with substantive significance by using methods that capture

whether a relationship matters (Schrodt 2014). If a variable's coefficient is non-zero and statistically significant, then it may be relevant, but by failing to consider how well a variable or variables predict the dependent variable, a useful measure of the strength of a relationship is ignored (Hindman 2015). By focusing solely on statistical significance, a useful set of alternative approaches to determining the veracity and strength of an association are abandoned. The predictive approach represents an effort to explore the strength and utility of an association, one which overcomes some of the problems of significance testing.

## Problems with Significance Testing

Prediction constitutes a means of testing the relevance of a variable or collection variables that differs from statistical significance, and is becoming necessary due to observed difficulties with significance tests. The problems with significance tests originate with the interpretation of their most common manifestation, the p-value.

The p-value is difficult enough to interpret that methods teachers regularly pass on incorrect interpretations of the statistic to their students (Haller and Krauss 2002). To remedy this, Haller and Krauss suggest that the solution to methodologists' difficulties in interpreting and teaching p-values lie in developing the logic of significance testing and comparing it to Bayesian statistics. They also argue that teaching the controversy around p-values is a useful way of communicating their true role to undergraduates, listing articles that detail complaints against significance testing. While discussing this controversy, Haller and Krauss advance it. Despite their attempt to stay out of controversies surrounding significance testing, however, Haller and Krauss display a major problem with the p-value and the approach that uses it, showing that its

lack of clarity represents a hindrance to advancing knowledge.

Despite arguing that the methods teachers failure to correctly interpret p-values represents a failure of pedagogy, rather than a problem with significance tests, Haller and Krauss elegantly demonstrate a need for a better alternative. While comparing alternatives and teaching how p-values have been challenged may represent a better way of teaching statistics, a viable alternative could be to develop a statistical logic that can be easily comprehended and taught. The predictive approach provides such a logic, and in conjunction with ideas like overall error rate (Kuhn and Johnson 2013), allows for a way of circumventing problems with significance testing by providing a logic that can be easily comprehended. This is in part necessary due to the difficulties in understanding significance testing, but also due to the problems that exist in its execution.

If the difficulties of significance testing lay exclusively in the pedagogy surrounding its teaching, then statistical significance might remain the best way to determine how two variables or collections of variables relate to one other. Statistical significance, however, is rife with difficulties, from to the logic that underlies the test (Gill 1999), to the possibility of false positives and false negatives (Verzani 2005), to the frequent misuse of significance testing (Gelman and Loken 2013). While the type 1 error may represent a means of explaining the likelihood that a hypothesis is rejected falsely (Verzani 2005), its presence means that significance testing remains easy to misuse or misinterpret. Such difficulties lead to problems with the way research is conducted; when significance is the final arbiter of a relationship between variables, its flaws make attempting to manipulate the test attractive and inevitable.

The problems with significance testing originate with the difficulty in interpreting it, and

run into deliberate and accidental manipulations when tests are conducted. The reliance on significance testing has led to deliberate attempts to find significant results through repeatedly running different regressions with varying collections and operationalizations of variables until the desired result is found. This process, which Simmons, Nelson and Simonsohn defined as researcher degrees of freedom (2011) might be mitigated by registering data and making the process of analysis more open (Reinhart 2015). These protections, however, may not prevent researchers from doing this accidentally as any situation where variables may be chosen can lead to a similar situation, even if the deliberate intent to deceive is not present (Gelman and Loken 2013). Through significance testing, hypotheses that are unlikely to be true can be found as true due to the unwitting manipulations and deliberate choices of researchers, which is troubling given the discipline's complete dependence on significance tests.

It is frequently possible to find statistical significance for tenuous relationships thanks to deliberate choices of operationalization and variable inclusion, even if the relationship is unlikely to reflect the truth of the hypothesis. With significance testing, variables that have little substantive relevance can be determined as statistically significant and are added to a massive list of relevant predictors (Schrodt 2014). The reliance on significance testing has led to problematic situations and conclusions, with individuals considering the independent variables' actual relevance secondary to the size of their p-value, a problem that directly focusing on how well variables' predictive capabilities should avoid.

A variable's ability to predict may be tested in a variety of ways, from information criteria to holdout sets (Clark 2004), but explicitly considering the ability to predict separate from explanation will improve analyses. Too often there is a discrepancy between how results are

analyzed and what the models and variables can explain or predict, due to the conflation of explanation and prediction. Predictive power cannot be inferred from explanatory statistics (Shmuelli 2010), and it is necessary to consider a separate set of tests and methods to determine predictive power. For these tests, a new approach is necessary, so that the ability to model and predict may be given full consideration.

While a predictive approach may be necessary as a way to avoid conflating two different statistical concepts, prediction and explanation, considering problems from a predictive perspective also presents a great opportunity for political scientists. By explicitly considering prediction separately, political science makes its findings more relevant to individuals operating outside the field (Sides 2014). This desire for more applied political science has benefits in allowing for a more interdisciplinary look at politics, and enables a certainty that was not there previously. Such a consideration cannot be the primary concern of the discipline, but the ability to effectively communicate results and to create relevance in a greater context should be a goal of political scientists, enabling a greater engagement with other disciplines and world at large.

Thanks to difficulties in interpretation and execution, there are difficulties in using significance testing alone to determine whether a variable's effect is relevant for predicting new observations. Although it has been dismissed as being about profit rather than understanding (Shmuelli 2010), prediction allows for a statement of substantive significance that statistical significance testing completely eschews. While the predictive approach bears similarities to effect sizes and confidence bands, which can indicate substantive significance in an explanatory framework (Cohen 1994), it differs in that it provides an additional test of the model, one focusing on its practicality. This has been dismissed by the literature (Shapiro 2002), but by

ignoring the possibilities that prediction holds for determining substantive relevance, we ignore a useful test of theories and models. In addition, prediction provides a stronger statement of relevance than the explanatory techniques that Cohen favors. By concentrating on prediction and its use through new techniques, the independent variable isn't simply related to current observations, but is demonstrably relevant through its ability to determine observations which have yet to be collected, surpassing statistical significance.

### But why BeSiVa?

Having discussed the utility of prediction as an approach that might break through some of the problems related to explanatory statistics, the use of my new algorithm, BeSiVa, rather than other approaches, must also be considered. After all, a regression could be run on a subset of data, and its ability to predict could be considered another set of data, without requiring an algorithm. Although this approach might work, it would divert from one of the main benefits of the algorithm. BeSiVa selects variables, excluding irrelevant variables from its conclusions, which a single regression cannot accomplish. Although other techniques that can also accomplish variable selection exist, BeSiVa is superior to these approaches due to the intuitive techniques it uses to make its decisions.

While BeSiVa may be able to select variables from a large list, other techniques are also capable of making such a judgment. These techniques include penalized regression approaches such as the lasso and elastic net penalized regressions, which are also capable of variable selection (James et al. 2008, Hastie, Tibshirani, and Friedman 2009). The BeSiVa algorithm, however, selects variables while creating logistic models, which is already used throughout the social sciences and whose principles are well known and understood. In other words, BeSiVa is

preferable to other techniques due to the addition of a feature, the ability to select variables, while using a technique with properties that are well understood by political scientists[1].

## What is the BeSiVa Algorithm?

### How the Algorithm Works

Once the data has been collected, recoded, and saved to a format that a programming language may understand, the algorithm requires three main inputs in order to operate. First, the algorithm needs the data. The data may have purely numeric, categorical, or numeric and categorical variables mixed together, but it must have a preset number of columns, where each column of data has the same number of rows. If the data can be read into a programming language from a spreadsheet, or a file from a programming environment like STATA or SPSS, it is capable of being used with the BeSiVa algorithm. Once the data has been read in, it is necessary to consider whether any of the variables need to be recoded. Although this depends on the question that a researcher seeks to answer, most variables do not need recoding for the algorithm to work, with the possible exception of the dependent variable. Once the data are read in, the only question is what to use with the Algorithm, and if any variables should be excluded.

While a majority of variables can be used with BeSiVa without incident, some variables

1. While the current use of OLS and GLM regressions are incapable of selecting variables like BeSiVa can, the possibility of using a technique like ensemble models might lead to better results overall. Ensemble methods, joining multiple types of model into a combined model whose predictions are superior (Siegel 2013), have the potential to create more predictive models in comparison to BeSiVa, which relies purely on logistic regression for its predictions. The results that arise from an ensemble method, however, are much harder to interpret, and the inner workings of such an approach have been described as a 'black box', with uninterpretable, highly complex prediction equations (Kuhn and Johnson 2014). By comparison, BeSiVa relies primarily on logistic regression, combining the predictive approach with a technique whose properties are known, making the algorithm's results comparable to other findings common to political science (Achen 2002). Although ensemble techniques may be capable of providing more accurate predictions, (although their implementation on real data makes this disputable, as seen in Rogers 2014), the BeSiVa algorithm's use of logistic regression leads to understandable models with accurate predictions.

need to be excluded from consideration. This includes any categorical variable with a large number of categories, as the risk of being unable to run a regression increases as the number of categories does, and it increases the challenge of interpreting the final model. The only limit, however, is the number of observations in the data that will be used for regression, which is the maximum on any technique that incorporates regression (James et al. 2013). Given that the data will be divided for different purposes, this will be slightly less than the total number of observations in the data, and will be the number of observations dedicated strictly to regression. It is also recommended that if a variable has a small number of categories, with one far more prevalent than the others, such a variable should be excluded from consideration. These variables are described as having low variance, and run the risk of being equivalent to having a variable with only one category (Kuhn and Johnson 2013). Since a selection of the data will be removed for testing purposes, such a variable also runs the risk of having no variation, making it inappropriate for regression. Apart from these potential pitfalls, the algorithm can use any variable that is provided as an independent variable, including those with missing data.

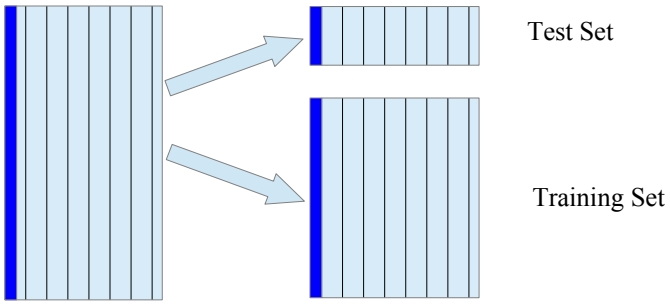
The algorithm is capable of dealing with missing data, as the regression commands that it uses delete any row with missing observations. However, it is worth checking the content of variables to verify that missing data is appropriately labeled and not an overwhelming proportion of the observations in the variable. Due to its use of logistic regression, the algorithm employs listwise deletion, removing rows where variables have missing data.

It is also necessary to make sure that the missing values are labeled correctly. Should a value be incorrectly labeled, the algorithm may fail to use the variable correctly if it is numeric, treating missing data inappropriately. It is also worth examining to determine whether a variable

is mostly missing data. The risk of errors increases when variables mostly including missing data are included, and it is recommended that such a variable either be excluded from consideration, or that the data be imputed. Once any recoding is complete, and missing variables are dealt with appropriately, the dataset is ready to use with the algorithm.

In addition to a well formatted dataset containing the independent and dependent variables, the algorithm needs to know which variables in the dataset will serve as the dependent variable, and the collection of independent variables that should be considered. Given the BeSiVa algorithm's use of logistic regression to create models, the dependent variable must be numeric and binary, with values of zero and one. In addition to the data and the dependent variable, BeSiVa requires a list of the independent variables' names. Once the data, and names of independent and dependent variables are provided, BeSiVa has sufficient information to run, choosing predictors that fit its criteria for most useful in modeling the binary dichotomous dependent variable.

Figure 1: an illustration of dividing the data. The data are divided to prevent overfitting, defined as when the whole of the data are modeled well, but outside observations can't be predicted.




The first step in the BeSiVa algorithm involves taking the dataset and dividing it into two separate subsets of data, as pictured in figure 1. The algorithm uses each of the datasets for different purposes, making such a division necessary. To create these data subsets, the algorithm

chooses a collection of observations at random. The algorithm selects one set of observations to test how well each model predicts the dependent variable, defined as the test set or test data. The algorithm estimates models using the remaining set of observations, which are referred to hereafter as the training set or training data. Once the division of data is complete BeSiVa begins estimating models using the training set, and determining their ability to predict the dependent variable using the test set.

Once the data has been divided, the algorithm models the dependent variable using a loop, which saves the results of the models and determines which model makes the best predictions. The main loop first makes the formulas for the models. In its first run, the loop takes every independent variable that the algorithm needs to consider, and creates a formula for regressing each independent variable against the dependent variable. Once the first iteration is complete, the algorithm will save the model that creates the best predictions, and save its independent variable for use in all future models. For the first run of the algorithm, however, each independent variable is put into a formula with the dependent variable, and is used to regress each independent variable against the dependent variable.

Once the formulas have been generated, the algorithm begins creating logistic models. These are generalized linear models, allowing for the estimation of logistic regressions. In the first iteration of the loop, the algorithm uses each of the formulas created in the previous step to estimate a logistic regression using only the data in the training set. In future iterations, these one variable models will be superseded by models that keep the independent variable or collection of independent variables that best predicted the dependent variable, but the first iteration makes one model for each independent variable. These models can then be compared by taking the test set

and using it to make predictions, determining which model makes the best predictions of the dependent variable.

Having generated a model for each variable in the selection of independent variables, the algorithm must now compare the models, and decide which model is most capable of predicting the dependent variable. These predictions, however, are only generated on the dependent variable in the test set, which has been kept away from the algorithm's model estimation. The predictions are then used to classify the dependent variable. The algorithm must decide, based on their most likely value, whether each observation in the test set is most likely to be zero or one.

Once the algorithm creates predictions for each model, those predictions must be compared against the measured values of the dependent variable in the test set. Determining which model creates the best predictions is difficult for logistic regression, and is done using a classification based on the predicted probabilities. If the measured value equals the most likely predicted value, then the prediction is determined to be correct, and is considered incorrect otherwise. Logistic regressions, however, do not generate simple binary outcomes, but probabilities (Rabe-Hesketh and Skrondal 2012), which must be compared with the test set's values of the dependent variable. To predict the outcome in a manner that can be compared to the test set, BeSiVa classifies each probability based on its most likely outcome.

To compare the model's results to the measured dependent variable, the algorithm decides which values of the dependent variable in the test set should be predicted as one or zero based on the probabilities predicted by each logistic regression. If a dependent variable's predicted probability is 50% or more, then BeSiVa classifies it as having an outcome of one, with any probability below that value classified as zero. This classification captures the most likely

outcome, without having to simulate a draw from a binomial distribution, which is how the dependent variable for the logistic regression is generated (Rabe-Hesketh and Skrondal 2012). This approach to predicting the dependent variable, however, guarantees that the algorithm's predictions are reliable, with the same probability leading the algorithm to make the same prediction each time.

The algorithm uses a classification scheme based on the most likely result of predicted probabilities, leading to a more reliable prediction than a draw from a binomial distribution. While there may be a more accurate probability to use as a cutoff, fifty percent demonstrates that no prediction outcome is more important than any other. If getting true positives is more important than true negatives, the probability cutoff for ones and zeros may be altered to better capture the more important outcomes (Kuhn and Johnson 2013). Since no prediction is more important, however, this approach guarantees a reliable, fair way of making classifications that predicts the dependent variable in the test set.

Having discussed how observations in the test set are predicted, it is important to consider what would happen in the presence of missing data. If a prediction can not be made (for example, in a row in the test data with missing data in one of the independent variables), then the algorithm treats the result as though it were incorrect. This decision penalizes independent variables with large amounts of missing data, allowing the algorithm to take the feasibility of using a particular variable in its models into account. These allow for the prediction or lack thereof of all observations in the test set, which can be compared to one another to select the most capable model.

Having generated a set of predictions for each model, the algorithm must now determine

how well these predictions were made. The algorithm determines prediction quality by dividing the number of correctly predicted values of the test set's dependent variable to total possible correct predictions, as seen in formula 1. Since the algorithm made these predictions for the test set, the total possible predictions are the number of rows in the test set. This can be thought of as a percentage, which is hereafter known as a model's percent correctly predicted (PCP for short) and is expressed in formula 1 below. The PCP is a way of determining a model's ability to make correct predictions, and is used in comparing models created by the algorithm, with one PCP created for each model that the algorithm estimates.

$$PCP \equiv \frac{number\ of\ correct\ predictions\ of\ the\ dependent\ variable\ in\ the\ test\ set}{number\ of\ observations\ in\ the\ test\ set} \qquad (1.)$$

Having generated a PCP for each model, the algorithm looks for the model with the largest PCP. This model's independent variable (or collection of independent variables in iterations after the first) are considered to be the independent variable(s) most capable of making a prediction of the dependent variable. The algorithm removes this independent variable or variables from future consideration in estimating models, as it will keep variables in the model with the highest PCP in all future models it makes. The highest PCP, however, is determined by the proportion of correct predictions to the total number of observations in the test set, to mitigate a bias towards variables that predict a small number of observations extremely well. This may be changed if the alternative is preferred, but the algorithm was designed initially so that all variables' predictions are equivalent. The determination of the model with the highest PCP represents the last step in the main loop of the algorithm, leading it to either repeat with the variables from the best model included, or stop if one of several conditions are met.

When checking to determine what model is most capable of predicting the test set,

BeSiVa makes a choice based on the PCPs of the models it has created. If one model has the highest PCP, BeSiVa keeps the variables in that model in all models in the next iteration. If there are no more iterations to run, however, the algorithm outputs the variables that led to the highest PCP. Once its operation has ceased, the algorithm provides the independent variables and PCP in the best model, among other outputs which allow the user to evaluate and easily create models based on the results of the algorithm. The algorithm, however, must also deal with the possibility of ties, that two models have the same PCP, something which it is not equipped to handle.

Since the algorithm has no rule designed to break a tie, its operation ceases if two observations have the same PCP. If more than one model is tied for the largest PCP, the algorithm stops, outputting the independent variables from the iteration before the tie occurred. While this may leave a potentially useful variable on the table, the algorithm is capable of selecting variables, but its predictive criterion, the PCP, leaves no way of determining which of two equally matched collections of variables is preferable. This allows the algorithm to make a unique selection of variables that maximize PCP, and avoids dealing with ties, which the algorithm is incapable of managing.

Once BeSiVa has finished running, it outputs results from the algorithms. First, the algorithm provides a list of independent variables that it deemed most useful in estimating models. The algorithm also outputs the list of tied formulae on the last run of the algorithm, if there was such a tie. Finally, it provides the formulae from the final run. It also provides the PCP values from each model in the final iteration, as well as the rows of data used in creating the test set. While a number of these elements are output for diagnostic reasons, the algorithm's results are meant to give researchers a chance to understand BeSiVa's motivations for choosing the

results that it did. It also allows researchers to decide whether it is necessary to shape or change the output through additional options.

To run BeSiVa, it only requires a dependent variable, list of independent variables, and the data, but a selection of options are provided to allow researchers to shape the results as desired. Perhaps most essentially, the algorithm allows researchers to change the number of iterations in its main loop. Since this loop determines the number of times the algorithm looks for variables in the provided data, this option effectively selects the maximum number of independent variables the algorithm can choose. It can not control the minimum number, as a tie will always break the algorithm's operation, but it lets the user decide how many times the algorithm can search at most.

In addition to allowing researchers to determine how many variables, at maximum, BeSiVa is allowed to find, the algorithm also offers the ability to change the size of the test set. The test set's size is reflected as a proportion that can range anywhere between zero to one. This argument allows the user to choose the percentage of data to hold out for testing purposes, with a default of 0.2, or twenty percent held out from the models to test. Although the percentage chosen for the test set is a tuning parameter, meaning that it lacks a mathematically optimal value, it is based on conventions related to k-fold cross-validation (Kuhn and Johnson 2013), upon which BeSiVa is based, and allows for experimentation based on the research question and the number of observations.

In addition to determining the proportion of the data used for testing, the algorithm also allows the user to add a threshold for the minimum allowable increase in the percent correctly predicted. This threshold rounds the PCPs, making it so that no increase below that threshold

counts as an improvement. This threshold's default value is set at a generous $1 \bullet 10^{-6}$, effectively stating that the minimum improvement considered is one ten thousandth of a percent, a permissive value for a variable's contribution. This threshold option gives a researcher the chance to fine-tune the algorithm's decision-making process by stating what improvement constitutes a meaningful contribution by a variable.

Determining the appropriate size of the improvement threshold for including a variable in the model depends on the problem. The scale of the threshold, however, is the same as the percent correctly predicted, requiring a value between zero and one. A researcher may decide that any improvement is sufficient to include a variable in the model, setting the threshold to 0, or they may choose a seven percent improvement by selecting 0.07 as a threshold. While there is no optimal value for this threshold, as it increases, BeSiVa becomes less likely to select more variables, as the threshold keeps it from choosing variables that improve the model below its value. This threshold lets a researcher state what increase in PCP represents a large enough improvement for a variable to be selected for inclusion in the algorithm's recommended list of independent variables.

In addition to controlling the algorithm's selection criteria, BeSiVa's options include the chance to control the random split in the data by changing the random seed. The random seed allows for control over the random number generator. This lets a researcher control how random the results are (Matloff 2011), which is useful for verifying the results of the algorithm. Since the algorithm divides the data at random, the final option lets a researcher repeat the same split for the purposes of replication. This allows a researcher to get the same results, as the same division of data will result in exactly the same selection of independent variables every time,

guaranteeing complete consistency in the results. This can be repeated with different seeds, allowing for a selection of the same results to be obtained for the purpose of Monte Carlo simulations. While the other options allowed for different ways of shaping the results, setting the random seed lets the user decide whether the same results should be obtained each time. This ability to get consistent results may seem counterintuitive for an algorithm that relies on random splits in data, but it is useful for getting consistent results in a single run or set of runs for testing purposes.

The benefit of repeated splits in the data and running Monte Carlo simulations using the splits lies in the possibility of capturing exactly how well the dependent variable is predicted by the provided independent variables over many iterations. This allows for a consideration of the quality of predictions that would not be possible otherwise. While the consistency of the selected variables will be discussed in greater detail later, it is possible to how the predictions may vary between runs. For instance, the BeSiVa algorithm's PCPs can not be normally distributed, as they are bounded between 100 and 0 (and it is well known that the normal distribution is bounded between infinity and negative infinty), but as seen in figures 3, 5, and 6 they do cluster around a central tendency. In addition, the variables selected by the algorithm will vary depending on how well they predict the dependent variable in each run. When one variables or a selection of variables are capable of predicting the dependent variable well, the algorithm is highly consistent. In this context, consistency means selecting the same variable or set of variables in an overwhelming majority of cases, as seen in figure 4, where the exactly the same variable was selected in 97 out of 100 runs. The results of the algorithm's predictions allow for the consideration of how well a prediction is made over repeated observations, and

demonstrate not only if predictions can be made, but the consistency of those predictions as well.

### On Overfitting, or Why it is Necessary to Divide the Data

Of the many rules that make up the algorithm, one aspect of the process stands out as particularly counterintuitive: the division of data into two subsets for different purposes. This is counterintuitive from the standpoint of explanatory statistics, due to the increase in the size of standard errors, and thus the decrease in confidence about whether a predictor is nonzero (Verzani 2005). Using prediction as a criterion, however, such a division is necessary. The usefulness of a given collection of variables is not based on statistical significance or other attempts to estimate the goodness of fit, but on the percent of observations correctly predicted in the test set. The reasoning behind predicting a set of data that was not used in estimating the models involves preventing overfitting.

In order to understand why overfitting is something that is important to prevent in the predictive approach, it is necessary to consider how the approach BeSiVa takes differs from an approach originating in inferential statistics. This approach, commonly used in the social sciences, has its basis in testing an overarching causal theory or set of causal theories, expected to apply to the population. A sample of the population is used to make inferences about a population, with relationships based on passage or failure of significance tests (Caldwell 2011). If the test is passed, then the explanation is considered valid.

The approach of inferential statistics relies on significance testing and goodness-of-fit measurements to test a causal explanation, and whether the relationship specified by the

explanation is likely to apply to a greater population of observations. By comparison, the predictive approach focuses on applying statistical techniques for the purpose of taking observations and guessing their values as accurately as possible (Shmuelli 2010). For this reason, a given model's ability to predict must be tested, and the most reliable test of whether a model is capable of predicting new observations involves attempting to make predictions on a separate set of observations.

Part of the reason for dividing the data and predicting data that was not used in regresion is focusing on the model's ability to predict, but its most important reason, especially for an algorithmic approach, is to prevent overfitting. Overfitting, defined as following noise in the data more closely than any existing relationship (James et al. 2013), can be prevented by an approach that focuses on predicting data that was not used in the regression, such as a test data set (Clark 2004). A single test on one dataset, such as correlation or statistical significance, runs the risk of overstating relationships between dependent and independent variables. An example of this was demonstrated by Cureton, who showed that when a new psychological measure was tested on the same data used to create it, its ability to predict new observations would be dramatically overstated by a correlation test (1950). Similarly, if the only measures of a relationship between dependent and independent variables arise from explanatory statistics, or even predictive statistics on a single dataset, the chance that the relationship holds with other data remains untested and can be unlikely.

While testing on a single dataset may prove problematic, statistical significance's flaw from a predictive approach lies in being unable to see whether the independent variables and model are overfitted. Overfitted models do an excellent job of predicting data used for modeling,

but poorly when making predictions (Kuhn and Johnson 2013). Overfitting often happens when too many independent variables are used to predict the dependent variable, risking that noise in the dependent variable is predicted well, but new observations are predicted poorly. To make overfitting less likely, the data should be divided for different purposes, as following the noise in data used for regression is impossible if a new set of data is used to test the regression's relationships. This is why BeSiVa divides the data into two sets at random, at the cost of removing data and increasing the size of the standard errors. Even if statistical significance is affected, the predictive approach requires that any model can predict new observations, ones that differ from the measurements used for estimation purposes.

### Challenges of Using the BeSiVa algorithm:

Despite the algorithm's variable selection and predictive capabilities, several caveats are necessary to make sure its operation goes smoothly. First, the data must be considered carefully. One of Harrell's major critiques of a precursor of the algorithm, which also selected variables, was that “it allows us to avoid thinking about the problem” (1996). While the use of a test set, focusing on predicting rather than explaining, enables the avoidance of some of the other problems of this approach, the algorithm's operation can lead to a similar situation. Faced with a large number of variables or a new problem, the algorithm's results may be used in place of any overarching theoretical explanations. Although it may be used this way in conjunction with an inductive approach, the algorithm's employment with a strong theoretical argument is capable of transforming the way and the amount of certainty with which research questions are answered.

The BeSiVa algorithm, in attempting to determine which variable or set of variables best

predict a dependent variable, allow for a radical reconsideration of how data are considered. For example, by attempting to select the variables that best predict the dependent variable, BeSiVa allows for an analysis of data and variables that previously would have overwhelmed an army of theorists. As an example, the GSS data used in the empirical section of this chapter includes over 650 columns, and the algorithm looks at the predictive capability of each column before adding or eliminating to in the final result. As opposed to starting from a small preordained collection of variables, BeSiVa allows for a consideration of the majority of available data. This allows the algorithm to approach preexisting research questions and new problems from a standpoint that enables the discovery of new relationships or furthers the validation of previous discoveries.

While the algorithmic approach allows for the consideration of more data, attempting to find relationships among all variables rather than a small selection of available data, it also guarantees that the relationships that are discovered matter. Using prediction as an arbiter makes sure that new data are considered thoroughly, making any relationships that are discovered more likely to be useful and  are more likely to be sustained in future testing. By explicitly focusing not on a single hypothesized relationship or set of relationships, but on the whole of the provided data, BeSiVa allows researchers to not only consider the existence of discovered relationships, as well as an instant replication assessment (Hindman 2015). This way of considering prediction suggests that relationships that are verified predictively are more likely to reoccur in future testing, making prediction a suitable tool for checking if work is likely to be verified.

Prediction is useful for determining which relationships are likely to replicate, and it is also preferable due to the failure of political science to test predictions up until this point. Political scientists rarely test whether or not the data are predicted well by a recommended

model, assuming that statistical significance is sufficient to say that the dependent variable is predicted (Schrodt 2014). By explicitly determining the quality of a model's predictions, research makes a more compelling statement about a pre-hypothesized relationship. If variation is predicted, rather than explained, then not only do the results become more likely to replicate (Hindman 2015), but they make a stronger statement about the uncertainty of a relationship, something statistical significance does not explain (Reinhart 2015).

In addition to the discovery of relationships that are newer, stronger, and more likely to replicate, the predictive approach enables researchers to begin answering new questions in a more efficient manner. By concentrating on prediction, researchers may focus on understanding dependent variables that theorists have yet to consider, allowing for a more holistic way of thinking about the problem. Such an approach allows for a new kind of conversation between methodologists and theorists. With this approach, existing theory may be tested using new reationships, and may be rejected or considered further in an inductive test of new data and research questions. Via the use of prediction, and algorithmic means of determining the existence and strength of relationships between variables, the process by which relationships are determined may be strengthened. Such an approach allows for a better understanding of not only whether a theoretical relationship should be considered trustworthy, but also finds new relationships for theorists and methodologists to ponder in tandem.

While considering the data and the theory behind the algorithm's findings, it is also necessary to determine whether regression is possible or appropriate. While BeSiVa is still capable of running even if a few of the regressions that it conducts fail, no safeguards exist if the provided data are inappropriate for regression. For instance, one of the problems that logistic

regression presents is the effect of perfect prediction on the coefficients. If the dependent variable and an independent variable have values that correlate perfectly, then the dependent variable is said to be perfectly predicted, and the regression's coefficients will be unstable (James et al. 2013). It is necessary then to consider whether any of the predictors are inappropriate for logistic regression. The algorithm's dependence on logit then, requires a willingness to deal with the limitations inherent in logistic regression.

The limitations of logistic regression cannot completely be avoided, but they may be minimized through careful preprocessing. BeSiVa operates most effectively when data of low variance is eliminated. Data of low variance includes variables with only a single category or a category that is overwhelmingly present in the data (Kuhn and Johnson 2013), or data that primarily consists of missing values. If much of the data are missing at random, it is recommended to impute, as filling in the missing values will allow these variables to be considered by the algorithm . BeSiVa, however, can deal with unimputed data, although its reliance on listwise deletion leads to a risk of biasing the results.

The decision to impute is one worth considering due to the opportunity to better use variables with missing data, but imputation is not without its drawbacks. On the one hand, there is the risk that imputation is used with inappropriate data, where missing values are systemic, as opposed to missing at random. If this is the case, then it is not to impute, and to listwise delete, removing rows with missing data from consideration. This leads, however, to the possibility that listwise deletion biases the results (Hastie Tibshirani, and Friedman 2009). Due to BeSiVa's use of listwise deletion, the results may be biased in two ways, in favor of variables that are more complete but do not predict the dependent variable well, and in favor of variables which are

incomplete. It is necessary to consider each case separately.

Due to its use of listwise deletion, the algorithm risks choosing variables that are more complete, but do not predict the dependent variable as well as some variables with large proportions of missing data. Despite the preference for more information, this is not necessarily a major difficulty if predictions are desired, as the variables are predicting the dependent variable as intended. This reflects a preference for variables which are more complete, easier to gather and therefore more likely to be useful in making a prediction. A larger problem occurs if an independent variable that is largely incomplete predicts a small subset well, and is better capable of predicting the dependent variable than the more complete independent variables.

While it would be preferable if all data were complete, the algorithm can use listwise deletion to deal with missing data, predicting the dependent variable with the data at hand. A larger problem occurs if variables that are mostly missing values are still more capable of predicting the dependent variable than variables that are not. This problem was dealt with by forcing the algorithm to include missing data in the denominator in formula 1, meaning that a missing result is considered the same as an incorrectly predicted one. Such an approach allows the algorithm to consider missing data, treating the inability to predict as equivalent to making an incorrect prediction. If the PCP ignores missing data in its denominator, then it risks biasing the data towards variables with many missing values, biasing the results in favor of these data as displayed in figure 2. Although missing data are a problem that may be solved with imputation, the risk that the data are not missing at random, and the fact that the algorithm may cope with missing data means that a dataset may be presented to BeSiVa without imputation and considered with the minimum amount of preprocessing.

Figure 2: Why the algorithm treats missing data as an incorrect prediction. Figure 0 displays 3 variables: variable 1, variable 2, and the dependent variable. In the figure, black boxes in a variable represent missing data, which the algorithm treats as being incorrect. If it did not treat the data as incorrect, then the results would be biased towards whatever values remain. In the picture, variable 2 would have a PCP of 12.5%, while variable 1 would have a PCP of 87.5% if missing data is considered to be different from a correct prediction. If the algorithm did not consider missing data, by ignoring missing values in calculating PCP, for instance, then variable 2 would have a PCP of 100%. While it is possible to treat variable 2 as 100% right, the algorithm will choose that variable over variable 1, biasing its calculations and all future iterations in favor of the mostly missing variable.

| Variable 1 | Variable 2 | Dependent Variable |
|---|---|---|
| | | |
| X | | X |
| X | | |
| | | |
| | | |
| X | | X |
| | | |
| X | X | X |

In addition to questions related to data that was kept out due to missing values, another question related to the way the algorithm uses data arises. Specifically, what is the right size for the test set? According to Kuhn and Johnson, the test set's size is subjective, depending on the number of rows in the data. The size of this data is a tuning parameter, a part of the algorithm that has no mathematically optimal answer (2013), although the algorithm's operation provides some suggestions. In BeSiVa, however, increasing the size of the test set decreases the amount of data used for modeling, as the two are kept separate. This means that at minimum, a majority of the data should be left in the training set, but exploring multiple sizes of test sets as a means to compare the percent correctly predicted is highly recommended when using BeSiVa.

## Differences Between BeSiVa and Explanatory Approaches

While the BeSiVa Algorithm can guide understanding of a research question, its results

differ from those that arise from explanatory statistics in some notable ways. For instance, the 'true' model is left by the wayside in favor of whichever variables are most predictive. This also means that independent variables that are necessary when explaining the dependent variable may actually weaken a model's ability to predict (Shmuelli 2010). With a predictive approach, however, it is possible to determine which of the variables are most relevant from a substantive perspective, answering a question that statistical significance is unable to consider.

While it may appear that the algorithm is leaving useful variables aside, the question of how useful these variables truly are, remains open. The algorithm allows for the creation of predictive models, and for the consideration of research questions, but there is room for additional predictors when conducting in-depth, exploratory research. This, however, is not to say that such predictors are always relevant, but the use of statistical significance has led to a failure to make better explanations in favor of repeating the same approach (Schrodt 2014). In conjunction with strong questions, BeSiVa can be used in a way that allows for conversations with theorists in both inductive and deductive modes of inquiry.

Despite the benefits of using the test set to determine a model's capacity for making predictions, rather than focusing on statistical significance, the approach that BeSiVa uses to make predictions has a potential drawback. The algorithm's focus on a single test dataset may share a problem with data mining, if the algorithm swaps one unrepresentative dataset, the one used to estimate the model, with another, the subset of data that makes up the test set. While random sampling from the data is used to create representative datasets, this may not be the case if the sample is biased, or if a sample is poorly drawn (Kuhn and Johnson 2013). By relying on a single set of test data to make predictions, the algorithm risks the possibility that the test data is

unrepresentative of the whole. For this reason, BeSiVa may be used once, but this is a limitation of that approach, which is why the algorithm was also designed for ease in repeated use and bootstrapping. By repeatedly running the algorithm with different test sets, the potential for unrepresentative data is minimized, as such data will be drowned out by the results of more representative data, which will occur more often.

### Falsification: Determining BeSiVa's Overall Performance

A major difficulty of testing an algorithmic approach lies in the fact that, unlike a limitation of preexisting theory, political science has no established strategies for testing and falsifying a new method. In this, there is a risk that any tests of a new method lack a strategy for falsifying, that the method cannot fail, but can only be failed by the data. Fortunately there are several possible means of circumventing this problem. The BeSiVa algorithm can be falsified if it is unable to find appropriate predictors, if its predictors are consistent, but not theoretically relevant, and if the results it provides predict the data in the test set well.

In determining the utility of the BeSiVa algorithm, it is necessary for BeSiVa, or any variable selection method, to choose relevant predictors from real data. The use of real data instead of simulated datasets for performance might seem counterintuitive, until the purpose of prediction is considered. While explanatory statistics and significance testing concentrate on capturing the true model, the questions predictive approach concentrates on whether a set of variables may lead to a good prediction. Through the use of real data, the algorithm may be evaluated on its own merits, rather than the merits of an approach with different priorities, explanatory statistics (Shmuelli 2010). This use of real data then allows for the creation of a

series of predictive models, if they exist, which can suggest a way of determining whether the BeSiVa algorithm can be useful to political scientists.

While it makes sense to test a predictive algorithm's performance and ability to make predictions on real data, the determination of what makes its relationships worth considering remains open. By purely focusing on relationships between variables, the algorithm runs the risk of leading to spurious correlations, and incorporating them into its conclusions. What prevents BeSiVa from picking up the prevalence of ice cream sales in modeling whether individuals will drown in a month, to use a classic example of spurious correlation? One way to deal with this is to consider its initial results in comparison to theory. If theory is tested by methods, then a second way of testing a new methodology is by comparing its results to pre-established theory.

In an area where the theory has been well developed, such as the choice to vote, a useful method should support a theoretical approach, fulfilling the promise of variable selection and removing irrelevant variables from consideration (James et al. 2008). In order for a new method, one which claims to select relevant variables, to be trustworthy, it must demonstrate its ability in an area where theory is well established, by selecting variables that correspond to theoretically developed causal mechanisms. Through the creation of predictive models, and finding variables that are theoretically relevant, a predictive method such as BeSiVa must demonstrate its ability to validate pre-existing theory, so that such a method may eventually build upon theory.

### Initial trial: The Choice to Vote in the GSS

This initial demonstration uses the GSS, concentrating on the choice to vote. This version of the GSS was taken in 2014, allowing consideration of the most recent election for which a

comprehensive survey was taken. The choice to vote was operationalized with the statement of whether someone voted or not in 2012, which was transformed to a binary dichotomous variable, with 1 indicating that the individual voted, and 0 indicating that they did not. Given that only rows with the dependent variable could be considered, there were 2,374 observations in total.

In attempting to model the choice to vote in the 2012 election, the algorithm needed a collection of independent variables to consider. It was given a list of independent variables that included theoretically specified predictors, such as an individual's education, operationalized twice, by number of years that they were in school and the last degree that the individual achieved. It also included variables such as race, party identification, and whether someone voted in 2008. The list of independent variables, however, also included predictors that did not necessarily have a strong theoretical connection to the dependent variable, such as how a person felt about their health, whether they were salaried or paid by the hour, and even their astrological sign. The full list of independent variables was long, and while an attempt was made to be inclusive, the variables were pared down, and included based on their ability to function with the algorithm.

The list of independent variables provided to the algorithm included a collection of theoretically relevant variables, as well as irrelevant ones, only removing variables out that proved incapable of being used by the algorithm. Given that the dependent variable was vote choice in 2012, any column that served as a restatement of this choice, such as who a respondent voted for in the presidential election, was eliminated from the algorithm's consideration. Similarly, variables that were unlikely to provide good predictions, such as those missing over 80% of their data, or any variable with near zero variance (Kuhn and Johnson 2013) were also

kept from estimation. These variables, the dependent variable, and a series of specified options were given to the algorithm to allow BeSiVa to determine what predicted the choice to vote in the GSS.

While the minimum that the algorithm may need to run the data, an independent and dependent variable, a few other options were specified to test the functionality of the algorithm. Ten percent of the data, selected at random, were used in the test set, leaving ninety percent of observations for model estimation. The algorithm used a threshold of 0.001 as a minimum, meaning that variables would need to improve the PCP by a minimum of one tenth of one percent to be included. The options were otherwise left at their default values, most notably five iterations, meaning that the algorithm could select a maximum of 5 independent variables. With these options, the algorithm proceeded to run and select the variables that it found to be most relevant to the choice to vote.

Table 1: Regression estimates for BeSiVa with and without prior vote.

| | First Attempt<br>Estimate<br>(S.E.) | Removed Prior Vote<br>Estimate<br>(S.E.) |
|---|---|---|
| (Intercept) | -3.442*** | 3.150*** |
| | (0.343) | (0.263) |
| Voted in 2008 | 3.949*** | . |
| | (0.150) | |
| Education (Years) | 0.128*** | . |
| | (0.025) | |
| Graduate Degree | . | 0.549* |
| | | (0.262) |
| High School Diploma | . | -1.063*** |
| | | (0.151) |
| Junior College Degree | . | -0.745*** |
| | | (0.223) |
| Less than High School | . | -2.083*** |
| | | (0.193) |
| Weak Republican | . | -1.281*** |
| | | (0.271) |
| Lean Republican | . | -1.478*** |
| | | (0.272) |
| Independent | . | -2.491*** |
| | | (0.250) |
| Lean Democrat | . | -1.651*** |
| | | (0.261) |
| Weak Democrat | . | -1.552*** |
| | | (0.256) |
| Strong Democrat | . | -0.015 |
| | | (0.284) |
| Other Party | . | -1.830*** |
| | | (0.376) |
| N | 2239 | 2348 |
| Deviance | 1449.296 | 2372.294 |
| $-2LLR(Model\chi^2)$ ** | 1242.991** | 507.421* |
| AIC | 1455.296 | 2396.294 |
| PCPs | 88.6% | 79.7% |

$*p \leq 0.05 ** p \leq 0.01 ***p \leq 0.001$

The algorithm was run with the 657 variables that are listed in Appendix A, and the first variables that the algorithm selected were included in table 1, column 1. This table is the same as a standard logistic regression table, and was generated using the whole of the data. It has one additional component, however, which is the row entitled PCP. This row features the results of the algorithm's predictive criterion, the percentage correctly predicted of the test set. The PCP shows what percent of the 10% of held out rows were correctly predicted by the algorithm. The most notable thing about the 656 variables included from the GSS is perhaps the fact that even

with that selection, the algorithm chose prior vote, whether an individual voted in 2008. Without any prompting, the algorithm selected a variable that has grounds in the literature (Brody and Sniderman 1977, Gerber, Green, and Shachar 2003). Then, it did so again in its second iteration, selecting education as a relevant predictor of the choice to vote, before stopping due to a tie in PCPs. Not only are these two variables theoretically relevant, their ability to predict a majority of the test set data can also be verified. By including these variables, 88.6% of observations were predicted correctly, as seen in the row in the table marked PCP.

Through the algorithm, it is clear that whether or not someone voted in 2008 was an incredibly relevant predictor of whether they voted in 2012. To further test BeSiVa's ability to select theoretically relevant variables, the algorithm was run a second time using the GSS data, but with the prior vote variable deliberately left out. In this case, the algorithm could select from any of the other 655 independent variables, just not whether the individual voted in the 2008 presidential election. The algorithm was run with the remaining variables, and the results may be seen in the second column of table 1, where 79.7% of observations in the test set were correctly predicted. These percentages suggest that the algorithm is capable of selecting theoretically specified predictors, generating models which may prove relevant for the question of who is likely to vote.

When BeSiVa was tasked with selecting predictors apart from prior vote, it considered another elements of an individuals' demographic makeup that the literature discussed as relevant to turnout decisions. Despite choosing a different operationalization of education, BeSiVa still focuses primarily on education as a predictor, akin to the first iteration when prior vote was included. However, it also selects the strength of party identification as a predictor, which like

prior vote, is designated by the literature as a theoretical predictor of turnout due to its creation of attachments to the outcome (Campbell et al. 1960). Even when prior vote is eliminated also appears that the algorithm is selecting theoretically relevant predictors, fulfilling another falsification condition in these two tests. It may be, however, that these results are a function of the specific test data set, and therefore must be examined more rigorously to ensure their replicability.

### Repetition: Determining Predictive Capability.

While it is possible that the results above are representative of what can be expected from BeSiVa, it is necessary to test the algorithm repeatedly. This ensures that the model discussed above was not an unrepeatable event, with attractive results and high PCPs only occurring due to an unrepresentative test set. For this reason, the results should be repeated more than once, to determine whether there is consistency between iterations of the algorithm. If it turns out that these results do not reoccur, then repeating the algorithm with a different test set should provide evidence that this is the case, showing that the algorithm is not capable of generating predictive models. Thus, through repeating the process of running the algorithm with the same variables, it is possible to determine whether BeSIVa is capable of consistently making good predictions.

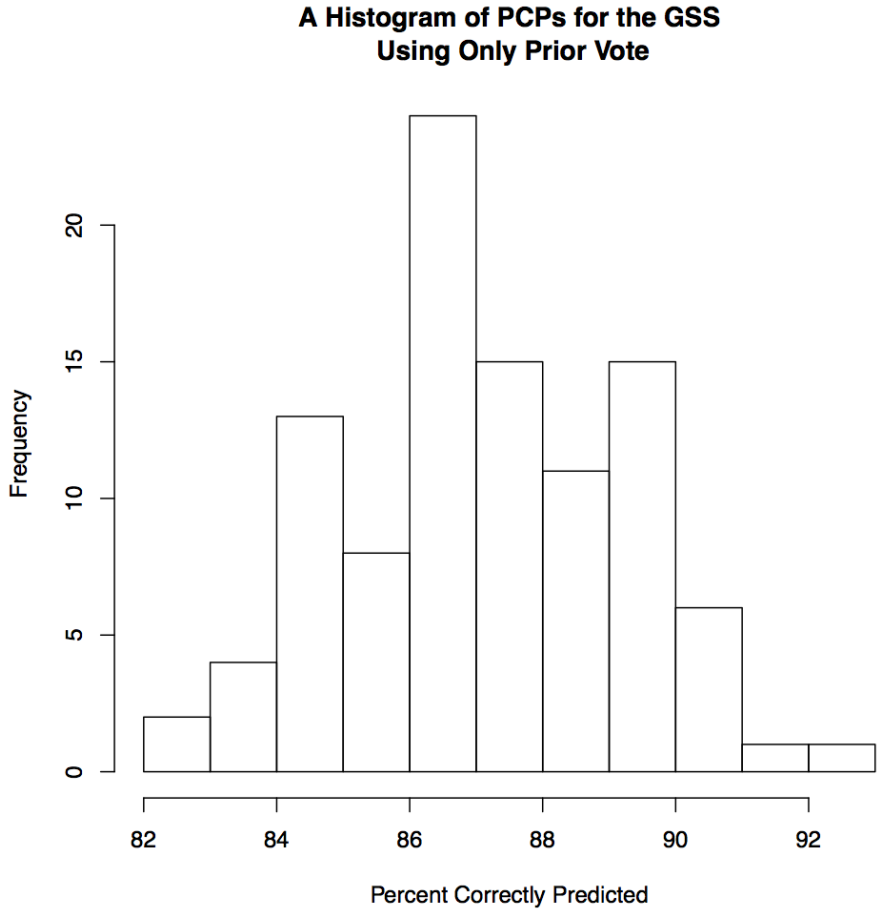
Changing the test set only requires rerunning the algorithm, changing which rows are selected for test and training sets. In doing so, it is possible to determine whether the results of a single run of the algorithm are representative, or whether the PCP and the predictors were a symptom of a test set that was different from the rest of the data. Given the fact that the algorithm needed to repeat the performance from the first run, it was possible to compare the

results of the initial test to other runs. The algorithm was run 100 times using a different test and training set each time, and the PCPs and variables selected were recorded. The results showed high predictions and consistent variable selection in the first iteration, with some instability in the second or third selected variable.

Table 2: Summary statistics for the percentages correctly predicted: The similarity between all variables and only prior vote suggests that it may be worth testing to see whether prior vote is capable of predicting the choice to vote entirely on its own.

| | All Variables | Removed Prior Vote | Only Prior Vote |
|---|---|---|---|
| 0% | 82.30 | 67.10 | 82.28 |
| 25% | 86.50 | 74.17 | 85.65 |
| 50% | 88.20 | 76.80 | 86.92 |
| 75% | 89.50 | 78.90 | 88.61 |
| 100% | 92.80 | 83.50 | 92.83 |
| Mean | 87.97 | 76.48 | 87.17 |
| Standard Deviation | 2.11 | 3.37 | 2.09 |
| Variance | 4.45 | 11.36 | 4.37 |
| Skewness | -0.12 | -0.16 | 0.10 |
| Kurtosis | -0.36 | -0.51 | -0.46 |
| NA's | 0.00 | 0.00 | 0.00 |
| N | 100.00 | 100.00 | 100.00 |

Figure 3: a histogram of the percents correctly of the 100 tests with all variables. When the 656 variables were included, the histogram was centered around 87.97% (with further summary statistics included in table 2), suggesting an excellent fit.



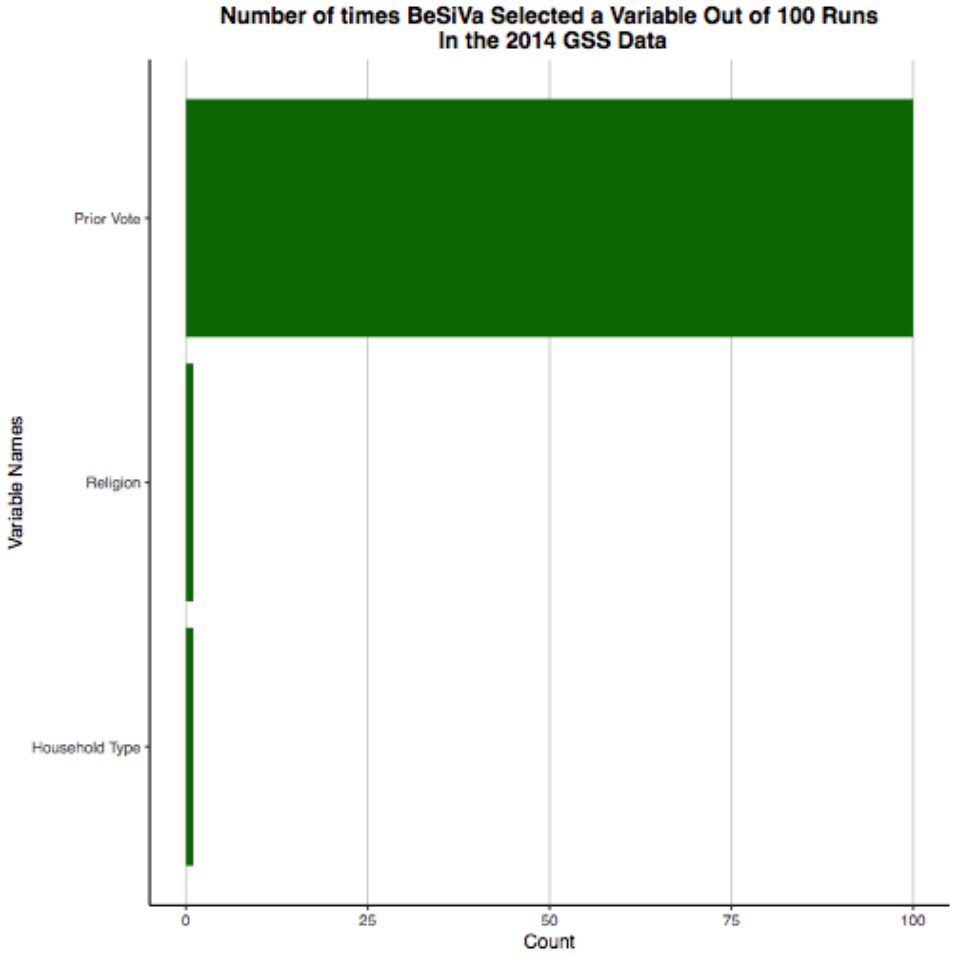

The algorithm was provided the same data, dependent variable, and list of independent variables that were provided in the initial single trial. The only thing that changed was the selections of data that served as test and training sets, dividing it differently each time. The algorithm only selected one variable consistently, but its ability to create predictive models remained consistent throughout the runs. This can be seen from the histogram and summary statistics of the PCPs, depicted in table 2 and figure 3.

Figure 3 shows the PCPs from the runs with all variables as a histogram. The x axis on

figure 3 displays percentages, showing that the PCPs for these runs are centered around 88%. The results appear normally distributed, but as percentages, they are theoretically bound between 0% and 100%. In practice, they also appear to be bound between 80% and 95% approximately, clustering tightly around a central value that suggests a high degree of predictive accuracy, compared to the threshold of one over the number of possible outcomes (Kuhn and Johnson 2013), or 50%. Depicting the PCPs from the run with all available variables, figure 3, demonstrates that given all available data, BeSiVa is capable of predicting the dependent variable with a high degree of accuracy

Figure 3 depicts the PCPs using a histogram, and table 2 examines the results of the 100 runs of the algorithm in more detail, providing summary statistics related to the PCPs gathered from each of these sets of runs. Each summarizes a set of runs with a different set of independent variables. The results may be thought of as the performance that can be expected from the algorithm when it is run many times on the same dataset with the same variables, and the first column depicts the PCPs for the runs from figure 3. As figure 3 depicted, the algorithm does an excellent job of making predictions when all variables are included, with a mean PCP of 87.97%, and a median PCP of 88.20%, as seen in the first column. The PCPs' maximum and minimum values are shown as well, and when all variables are included, the algorithm's poorest run is 78.90%, and it predicted the test set with 92% accuracy in its best run. Using the provided variables, the BeSiVa algorithm was capable of taking provided data and generating a predictive model, fulfilling one of the falsification conditions repeatedly.

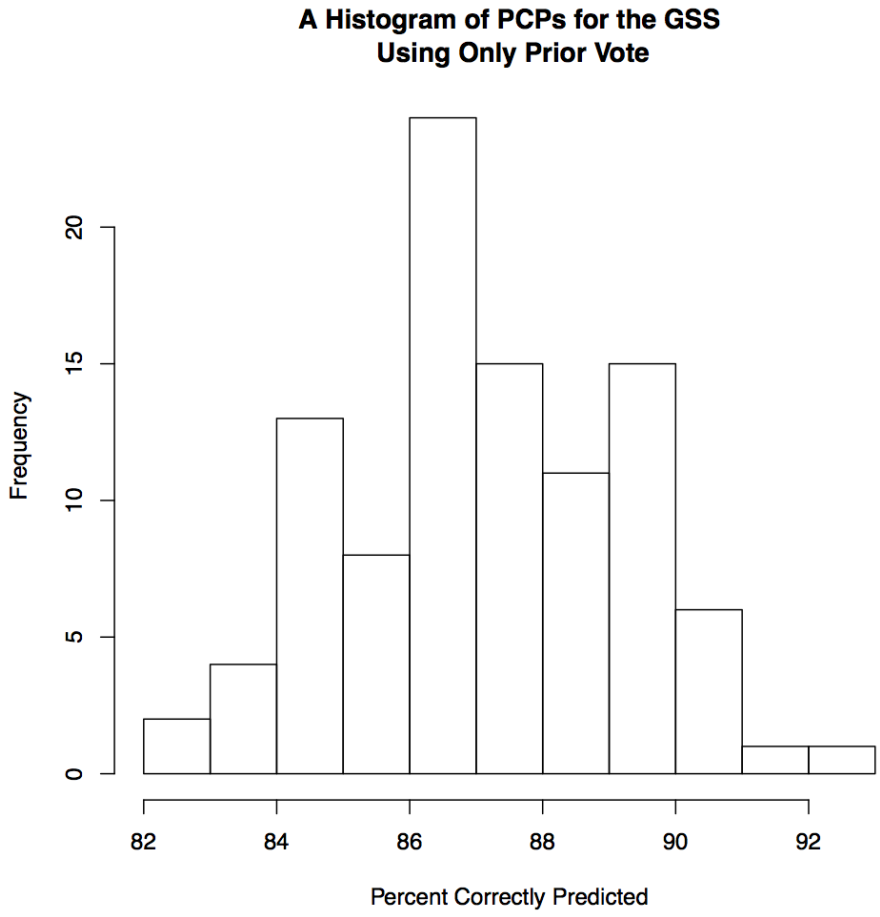
Figure 4: A bar plot of variables selected by the algorithm, when all variables were included. This plot shows any variable that was selected more than once by the algorithm. Once prior vote was selected, no other variable was necessary for predicting vote in the 2012 election in 97% of cases.




The algorithm may have fulfilled one of its falsification conditions, creating highly predictive models, but it is still necessary to consider the variables that the algorithm selected. It is clear from the results that there is little instability when the algorithm is allowed to select freely from the available variables, as seen in figure 4, a bar plot of how often each selected variable was chosen. While whether someone voted in 2008 was selected every time the algorithm was run, the question of whether education is necessary is called into question by these findings. Here, the second variable selected is the number of type of household that a person

grew up in (such as single parent, nuclear family, or other familial types), with religious belief following closely behind. There are so many variables selected, however, that it is unclear whether the algorithm's secondary suggestion of education is truly worth including in any model to make a good prediction, and demonstrating some instability in the process.

Figure 5: A bar plot of variables selected by the algorithm, when only prior vote was included.




One way to determine whether education is truly necessary in generating the most predictive model involves considering how well a logistic model that only included prior vote predicts current vote. This was tested with a bootstrapping process, with results displayed using

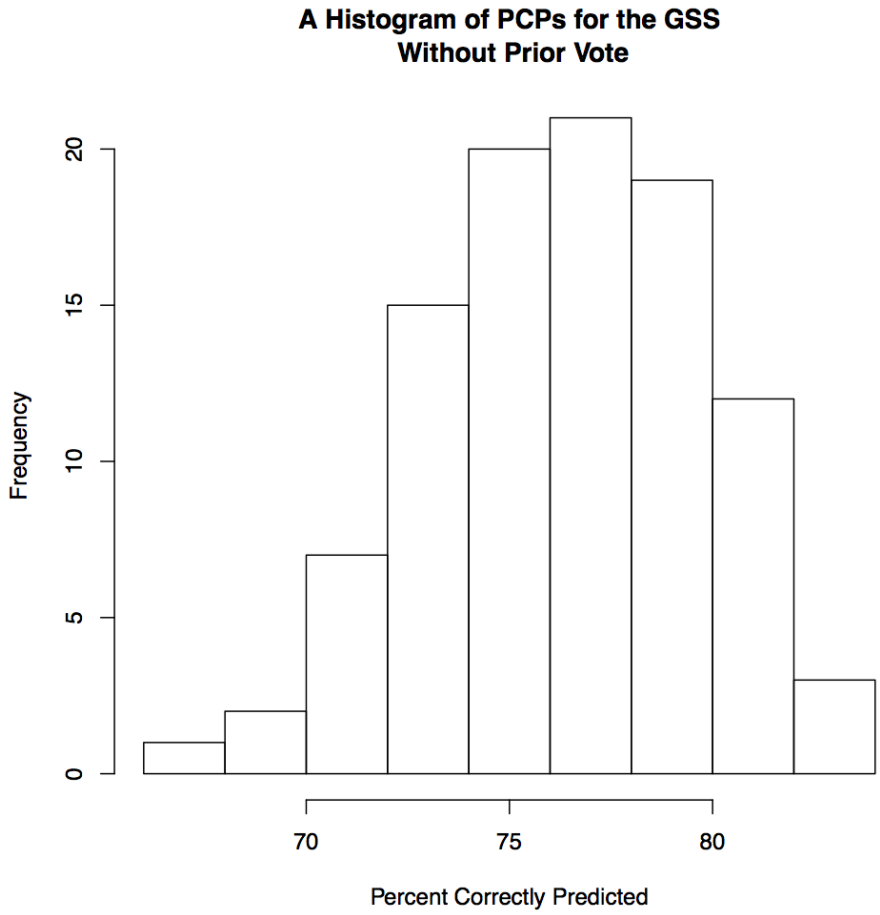
the histogram in figure 5, and the third column in table 2. To test the predictions, a process similar to the BeSiVa algorithm was used, except keeping the same variable in all regressions at all times. Ten percent of the data were kept out of the regression for estimation purposes, but the regression always had a dependent variable of voting in 2012, and an independent variable of prior vote in 2008. The results of this test demonstrate that when only prior vote is included, the PCPs are centered at 87.17%, similar to the results of the algorithm with other predictors. Even without education or any other independent variables, prior vote serves as an excellent predictor of whether someone voted in 2012. The question then becomes how to determine whether any other independent variables are necessary to improve the prediction.

Given the similarity to the results of BeSiVa, and the fact that it selected prior vote in every run, the question of whether another independent variable is needed when prior vote is accounted for remains open. One way of testing this involves calculating confidence intervals around the mean, an approach recommended by Cohen (1994) as a replacement for significance testing. Although Cohen was suggesting confidence intervals to describe effect sizes, the overlap of confidence intervals demonstrate that no additional predictors are necessary, even if they do improve the fit by a small amount. These confidence intervals were calculated and may be seen in table 3. Akin to confidence intervals on coefficients or means, these values show what the bounds on the PCPs are likely to be in a large number of repeated runs with the different sets of variables that are compared. The confidence intervals of the PCPs' means of just prior vote and the runs of BeSiVa that included prior vote overlapped, suggesting that although the algorithm regularly suggested more than one predictor, only prior vote is truly necessary to create a predictive model of whether someone voted in 2008.

Table 3: Confidence intervals on the percents correctly predicted. When making a predictive model, only prior vote and all variables have overlapping bounds, suggesting that only prior vote is necessary if a predictive model is desired.

| | Lower Bound | Upper Bound |
|---|---|---|
| All Variables | 87.55 | 88.38 |
| No Prior Vote | 75.81 | 77.15 |
| Only Prior Vote | 86.76 | 87.59 |

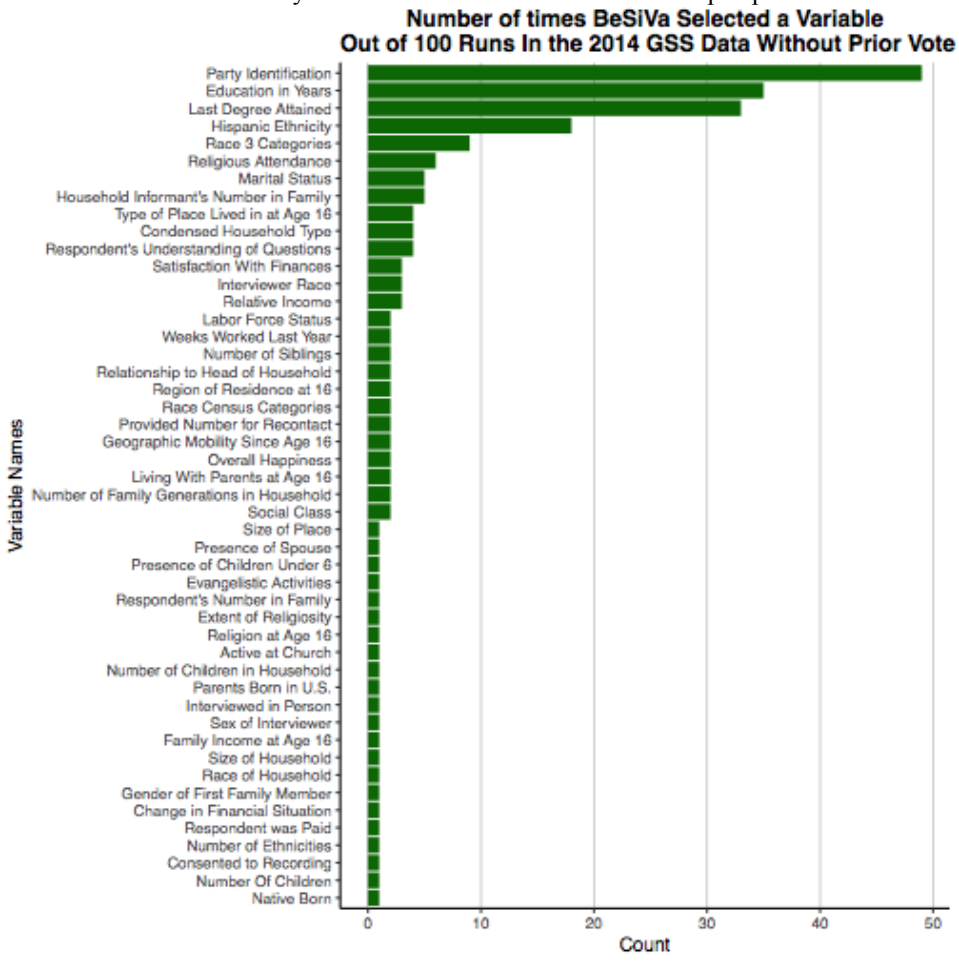
Figure 6: What happens when prior vote is removed. Without prior vote, the data are slightly less symmetric, and are centered at a lower value around 76.48%




Although it appears that only prior vote is necessary for creating a predictive model, it also may be that prior vote is not unique, and comparable results may be gained from a model

with the right set of variables. If this is the case, the algorithm would still be able to create or suggest models that have similar PCPs with the runs and model that included prior vote. To test this, the algorithm was rerun 100 more times using a different set of available independent variables. All variables except for voting in 2008 were provided, and the results of these runs may be seen in figure 6 and tables 2 and 3 under the columns that eschew prior vote. Meanwhile, Figure seven shows that the algorithm is less decisive than when prior vote was included, as no variable is selected with the same consistency as prior vote in figure 4. Party identification may be the most selected out of all possible independent variables, but it is only selected 49 times in comparison to prior vote, which was selected every time the algorithm ran. Education, meanwhile, served as the second and third most selected variable, due to differing operationalizations of the IV, immediately before questions related to race. While the variables selected by BeSIVa remain theoretically relevant, removing prior vote as an option creates an indecisiveness on the algorithm's part, suggesting that no combination of variables may replicate prior vote's predictive capabilities.

Figure 7: The bar plot when prior vote is removed. Once again, the plot displays any variable that appeared more than one time. Without prior vote, the algorithm becomes less decisive, never selecting a single variable in over 50 of the runs, but the ones it does select may be more attractive from a theoretical perspective.




While the algorithm still selects variables that are theoretically relevant (Campbell et al 1960, Teixeira 1987), when prior vote is eliminated, the BeSiVa algorithm does not select any of the variables as regularly as prior vote. This suggests that there is no combination of variables that can compare to prior vote for predictive ability and the decrease in the PCPs confirms the lowered predictive power of the remaining variables. In figure 6, the PCPs for the runs without prior vote may be seen, and the results suggest that a drop in the algorithm's ability to predict, as the PCPs cluster around 77%. This is verified by the summary statistics; the average percentage

correctly predicted for runs of the algorithm without prior vote is 76.48%, a drop of 11%, as seen in table 2. The confidence intervals for PCPs in the algorithm with and without prior vote, depicted in table 3 do not overlap, cementing the idea that prior vote is necessary to create the best prediction of whether an individual votes, and may be the only necessary variable for predicting the choice to vote well. While not necessarily as interpretable as some of the other predictors, it is clear that including prior vote alone generates more predictive models than all other independent variables if it were not included, although the theoretical implications of this variable require further consideration.

## Conclusion

There is a difficulty in the literature, concentrating on the lack of effort to make predictions. In some cases, prediction is dismissed as insufficient (Shapiro 2008), and it is considered reasonable to avoid testing or making predictions at all, even as a supplemental perspective. Too often, the literature makes no effort to predict the dependent variable, instead assuming that explanatory statistics are sufficient for the purposes of prediction (Shmuelli 2010). This has led to a situation which Schrodt decries as pre-scientific, where explanatory statistics, no matter how marginal, are considered sufficient to show a relationship and predict the dependent variable (2014). This situation suggests a need for an explicit consideration of prediction, one which can easily show if a proposed variable or explanation is relevant in considering a dependent variable.

An algorithm is proposed to make the best possible prediction on any data that are provided. The Best Subset in Validation algorithm, BeSIVa, makes predictions by creating a set

of logistic regression models on a subset of all provided data , starting with one model for each independent variable in this subset, known as the training data. It then tests how well each model predicts a separate subset of the data, called the test set. The test set was not used when algorithm created its models, preventing overfitting, a situation where data in the model are predicted very well, but the model is incapable of predicting data that wasn't used in the regression well (Clark 2004, Kuhn and Johnson 2013). The algorithm tests how well its models make predictions by generating a percentage correctly predicted, the percent of times the model correctly guesses the value of the dependent variable in the test set. It then keeps the model that makes the best prediction, and adds each remaining independent variable, repeating the process in an attempt to maximize the percent correctly predicted. This algorithm then provides the ability to consider how well any included variables predict a given dependent variable, which variables do the best job of making that prediction.

Having discussed the algorithm, it is then tested using real data to determine its ability to make predictive models and select relevant independent variables. In the first effort to determine BeSiVa's ability to predict, it is used on a selection of the 2014 General Social Survey, concentrating on the choice to vote in the 2012 Presidential election. In doing so, a large number of possible independent variables were provided BeSiVa for its consideration. The algorithm modeled whether someone was likely to vote, and stated that the best predictor of whether someone chose to vote or not was whether they voted in the 2008 presidential election, followed by one of the measures of their education. BeSiVa selected a model that made excellent predictions with theoretically relevant variables, which was tested further using repeated testing and varying which independent variables were provided to the algorithm.

In its first run, the BeSiVa algorithm suggested models that included two predictors and a highly accurate prediction of the data in the test set. With education and whether someone voted in the last presidential election, BeSiVa's model had a PCP of 88.0%, suggesting that the algorithm was capable of using the variables provided to make predictive models. BeSiVa even selected variables that mattered, despite a large quantity of irrelevant variables that were provided. This was repeated without the decision to vote in the last presidential election, which led to less accurate predictions. Despite this loss in accuracy, the algorithm was still able to predict whether a person was likely to turn out in 2012 with an accuracy of 79.7%, suggesting that while not as capable as a model with prior vote, a mix of party identification and education still created a predictive model. These theoretically relevant predictors and high levels of accuracy suggested that the algorithm had demonstrated its ability to find predictive theoretically relevant models with real data, avoiding its falsification conditions.

Despite the algorithm's ability to create predictive models, it was important to determine whether these findings were representative of what the algorithm might return if the data used to make predictions differed. To determine how the predictions might vary, the algorithm was run 100 times, using different splits between test and training data each time. While there was some instability after selecting prior vote, it was clear that the algorithm was capable of making predictive models, and that there was stability in its initial choice of variables. This situation continued even after prior vote was eliminated from the list of possible variables; the algorithm selected education and party identification, two highly theoretically relevant predictors, most often. This fulfilled the second falsification condition, that the algorithm selected theoretically relevant independent variables, which was fulfilled even after the most predictive variable was

removed.

To fully demonstrate the primacy of prior vote, the results using this variable were compared to other lists of independent variables provided to the algorithm. A set of PCPs were generated from models that only included prior vote, and were compared against confidence intervals generated by the full list of variables, as well as a list that excluded voting in 2008. From these confidence intervals, it is clear that prior vote would be sufficient to create a predictive model of current vote choice, and that its inclusion creates a superior model in comparison to all other possible variables. Although there was some instability, as BeSiVa selected variables beyond prior vote, it was clear that the algorithm was capable of finding the most relevant predictor from a theoretical and predictive standpoint. The algorithm selected a variable that had been discussed within the literature without any guidance apart from the predictive criterion and the resulting models made predictions that were centered at 88% accuracy.

The BeSiVa algorithm, when provided with a selection of data, was capable of fulfilling every falsification condition, including the creation of models that predicted the dependent variable, and the selection of relevant variables. It appears that if the only desire is to create a predictive model of vote choice, only prior vote is necessary to make the best predictions of whether someone will vote in the next election. Despite this, the question of whether prior vote truly is sufficient, especially from a social scientific perspective, remains open. The BeSiVa algorithm, however, is capable of providing variables and predictions that prove relevant not only from a predictive standpoint, but also from a theoretical standpoint as well.

# Chapter 2: Theoretical Utility and the Deductive Approach, or Demonstrating BeSiVa with Different Theories of Turnout.

## Introduction

The previous chapter focused on demonstrating how BeSiVa worked and what could be expected from its use, concentrating on what the algorithm found when it was applied to the General Social Survey and the choice to vote in the 2012 presidential election. The chapter included an exact discussion of how the algorithm worked, allowing for its recreation as desired. Once the explanation of the algorithm was complete, the algorithm demonstrated its utility on data related to the choice to vote. The results using BeSiVa showed that the algorithm is not only capable of making predictive models, but consistently predictive models. Most notably, when the favored predictor of binary vote in the previous presidential election was provided, BeSiVa chose only that variable 97% of the time. Although its consistency and the variables it selected demonstrate that the algorithm cuts away at enormous lists of variables, selecting useful predictors in the process, the question of how this may be applied in a manner consistent with social scientific practice remains open. This chapter focuses on using BeSiVa deductively, understanding the choice to vote using a collection of theoretically specified variables. In doing so, four relevant theoretical perspectives are considered, and three are tested and compared using the BeSiVa algorithm.

Having demonstrated the algorithm's capabilities at predicting a dependent variable with an eye for inductive reasoning, this chapter turns to how BeSiVa can be applied to voter turnout

and what a predictive approach may contribute when used deductively. The choice to vote has been considered from many different theoretical angles, with each one posing a differing explanation for why a person decides to turn out to vote. Through the algorithm, the different explanations, and therefore the reasons for including each independent variable, are weighed in comparison to one another, and a set of models using the algorithm's most commonly recommended variables are generated. These models are compared to theoretically specified models, each of which are tested using the same bootstrapping technique to determine how well they make predictions.

The BeSiVa Algorithm's ability to create useful models is based on the principles of predictive analytics, but it is used deductively here as an alternative to null hypothesis significance testing. Too often, the final arbiter for whether a variable matters or not is based entirely on its p-value (Schrodt 2014, Reinhart 2015), which explains whether a variable is statistically significant. Statistical significance, however focuses purely on whether a variable's true value is zero, saying nothing about the substantive significance of that variable (Cohen 1994). Using statistical significance as the deciding factor in whether an independent variable is truly relevant leads to questionable research findings and outcomes that only appear to make advances in our understanding.

The concentration on statistical significance has led to a series of perverse incentives for researchers that lessen the quality of research and the utility of its findings, calling seemingly settled conclusions into question. These questions are raised for a variety of reasons, including when an individual seeking significance inadvertently or deliberately creates models that

minimize p-values, even if the independent variables aren't necessarily relevant from a substantive perspective (Gelman and Loken 2013). There is also the problem of overstuffing models with an excess of independent variables, creating 'garbage can' models that show marginal significance for a variable of interest, violating parsimony (Schrodt 2014). Schrodt specifically points out that garbage can models trample on Achen's rule of three, which suggests that any model with more than three independent variables is incorrectly specified (Achen 2002). Researchers' focus on null hypothesis significance testing (hereafter NHST) as the final arbiter of truth has led to incorrectly validated hypotheses and models that are not parsimonious.

The models NHST incentivizes might be reasonable if they spoke to the utility of the theories they pretended to test, but this is highly unlikely given the way that the models are created and implemented. Model utility lies in the ability to not only explain the data that are provided, but to predict new observations as well. Prediction provides a separate test for these models, a test which takes the research further down the trail of inference (Mosteller and Tukey 1977). Prediction informs researchers about the overall utility of their models, how useful they are in a more rigorous test of quality than statistical significance. Imagine a voter turnout model that achieves significance from an explanatory perspective, validating the hypotheses that it sets out to test. But at the same time, the turnout model is incapable of predicting new observations. Even if the hypotheses were demonstrated yielded statistical significance, a model of whether an individual turns out to vote would be useless if it could not reliably determine if a person whose data was not used in the model was likely to vote or not. Unlike statistical significance, prediction explicitly considers how the model performs on new voters. In this situation, prediction serves as a useful test of models and hypotheses, leading to results that are robust to

repeated testing, making replication far more likely overall (Hindman 2015) In comparison to NHST, making predictions serves as a useful test of the substantive significance of models, and leads to conclusions that are likely to hold up in the long run.

While making predictions leads to a more rigorous test of hypotheses, failure to make predictions leads to results that are unlikely to replicate, ignoring a useful test of results and creating models that only appear to fit the data well. Attempting to capture all variation in a dependent variable, models tested via NHST may explain provided data well, but are also likely to overfit if new observations are considered. Imagine a model that, when data were provided, yielded the correct value of the dependent variable for every observation used in its creation. In doing so, however, this model failed to predict any observation that wasn't included in the initial regression. This is the problem of overfitting, when a model is good at giving accurate results for data used to create the model, and incapable of giving an accurate value for data that are not included Using all of the data for estimation then makes it difficult to determine if a given model is useful, as overfitted models may predict provided values of the dependent variable well and fail to predict observations that were not included in estimation (Kuhn and Johnson 2013). The predictive perspective needs to be considered for several reasons including parsimony, theoretical utility, and its general usefulness for testing hypotheses, making an approach like BeSiVa useful for reconsidering hypotheses.

But in light of the utility of the predictive approach, how can a technique like BeSiVa be used in social scientific practice? Using a deductive approach, the BeSiVa algorithm is here used as a means of comparing several different theories of voter turnout. This is in contrast to multiple

regression with NHST, which may allow for multiple theories' consideration, but does not choose between them, while NHST has systemic difficulties determining theoretical utility. Multiple theoretical approaches to turnout are developed briefly, before the variables that they specify to predict turnout are provided to the algorithm for testing. Once the algorithm has determined which variables are useful, the selected independent variables are added to models in order of their relevance, and these models tested using a separate bootstrapping approach.

The need for a better way to determining whether a theory is not only true, but useful, is called for. This chapter concentrates on using BeSiVa to make that determination among a selection of theories taken from the literature on voter turnout. It is determined that of these theories, psychological, sociological, and mobilization-based explanations, the BeSiVa algorithm's predictions favor the sociological and psychological explanations. Elements of these theories are found to be most useful for making a prediction, especially education, the strength of party identification from psychology, as well an individual's age, and the time that someone lives in a specific location. The comparison allows for a better understanding not only of what drives an individual to vote, but of prosaic elements that are too often ignored and suggest new directions for research on voter turnout.

## A Short History of Voter Turnout

### Psychological Origins and The Michigan School

The choice to vote has a lengthy history within political science, and the multiple causal

mechanisms that social scientists suggest drive the decision to turn out makes the area an ideal candidate for using a predictive algorithm deductively. Scholars have debated why people vote for over fifty years, leading to a variety of different theoretical models which could use comparison, beginning with the Michigan School. The founders of the Michigan school determined that individuals are more likely not only to make a choice in voting, but to vote in the first place due to their affinity for a political party. Party identification served as an indicator of interest, knowledge, and concern for the outcome of contests where the party a voter identified with was involved (Campbell et al. 1960). Campbell et al.'s original study of who individuals vote for, focusing on the role of party identification, provided a starting point for decades of research on the question of why people turn out to vote.

## Sociological Theories

Although they provided a strong explanation of why a person might decide to vote, Campbell et al. admitted to a potential lack of primacy in their causal reasoning, suggesting that no single variable could explain all aspects of the choice to vote. For this reason, voter turnout could be described as governed by other elements of a potential voter's life, a possibility first examined by Berelson, Lazarsfeld and McPhee. In their consideration of Elmira, New York, Berelson et al. concentrated on why individuals choose to vote in a specific way, but they also spoke to the reasons for turning out as well. Noting that "[n]onvoting is related to persistent social conditions having little to do with the candidates or issues of the moment" (32), Berelson et al. discuss the differentiation between social groups and demographics, and were perhaps the first to note the relationship between education, social group involvement, and turnout (1954).

Similar to Campbell et al. (1960) in their consideration of vote choice, Berelson et al.'s nascent sociological approach demonstrated a demographic component to explain why individuals turn out to vote, suggesting a new driver for political participation.

Although Berelson et al. considered the role of demographics in turnout, it was not their primary concentration in *Voting* (1954), and they shied away from explicit considerations of demographics in turnout. Instead, Berelson et al. concentrated on understanding vote choice through qualitative approaches, briefly considering why demographics might affect turnout. The decision to vote or not was examined further in later works, such as *Who Votes* by Wolfinger and Rosenstone. Using the United States Census' current population survey data, Wolfinger and Rosenstone posited that demographics -especially demographics that indicated resources that a voter possessed- could serve as a possible driver of the decision to turn out. Their findings suggested that from the sociological perspective, demographics that indicate resources are key to determining whether an individual is likely to vote.

Wolfinger and Rosenstone demonstrated the sociological approach by comparing a variety of demographic trends. In doing so, they showed that contrary to the class based suggestions of earlier researchers (Schattschneider 1975), the key resources to turning out were based in the abilities those demographics yielded for navigating bureaucracy and understanding voting. The role of these resources was compared along with the instrumental, expressive, and interest based benefits of voting, and the authors found that certain resource-based demographics mattered more than others. Due to its role as a provider of the ability to work with government, education mattered dramatically, imparting skills and experience necessary to maneuver a

government bureaucracy with ease. Similarly, age could impart the ability to navigate political situations, providing the experience to act politically even if a person had little formal education. In addition, whether someone was married mattered; marriage provided a strong interpersonal pressure to vote. In a paradoxical finding, an individual's wealth and free time, resources in the more traditional sense of the word, were not particularly indicative of a person's proclivity to vote. Wealth only mattered if the potential voter was unable to attain a level of comfort, and an individual's free time did not matter at all. If a person's occupation affected likelihood of turnout, it was only in terms of their interaction with government due to that job, making farmers more likely voters than other similar people. The key question then, in Wolfinger and Rosenstone's theoretical formation, was whether an individual had necessary resources, especially knowledge and experience, in order to vote (1980). While this element of cognitive resources is demonstrated, and their use of census data is groundbreaking, Wolfinger and Rosenstone were also limited by their reliance on the Census. The use of census data makes it difficult to consider alternative hypotheses of turnout, a limitation which hampers Wolfinger and Rosenstone's contribution to understanding political participation.

While the authors of the Michigan school were willing to admit their limitations, suggesting roles for life experiences, Wolfinger and Rosenstone's reliance on census data makes it difficult to control for alternative hypotheses of turnout, such as party identification's role in increasing the likelihood of turning out to vote (Campbell et al. 1960). This inability to test alternatives leads to a failure to consider partisanship, limiting the utility of *Who Votes* due to its inability to test theoretically established alternative hypotheses. Wolfinger and Rosenstone perceive the contribution of the Michigan school, but they are unable to control for its findings in

their own research. In addition, Wolfinger and Rosenstone mention a collection of causal mechanisms, including the roles of possible drivers of turnout such as a person's interest as well as instrumental and expressive benefits, but they fail to include any of those potential causal mechanisms beyond acknowledging their existence. For these reasons, Wolfinger and Rosenstone branch out from work that came before, but the limitations of their data makes it impossible to control for non-demographic alternatives, making difficult to describe *Who Votes* as building on prior research.

While they posit an alternative portrait of voting to the Michigan school, Wolfinger and Rosenstone are limited by their data, a problem which required a reconsideration of voter turnout. In his attempt to determine why voting declined in the aggregate, Teixeira focuses on individual turnout, and in doing so manages to overcome some of the limitations that hampered Wolfinger and Rosenstone. Exploring the findings of prior research while also further testing the sociological perspective's implications, Teixeira describes how links to party, the state, and media declined between 1964-1980, leading to turnout's overall decline. Despite this focus on the aggregate, Teixeira created an expansive theoretical picture of individual turnout, one which overcame the difficulties of the Wolfinger and Rosenstone's findings while illustrating a new difficulty of the turnout literature.

To Teixeira, each links between an individual and overarching institutions, the parties, state, and media represented a connection to different aspects of the political process. Links to party led individuals to have an interpretive framework for the issues, and made election outcomes personal. Links to media similarly gave an interpretive framework to the issues and

provided independent meaning to both the election and the issues a person cared about. And links to the state gave individuals motivation to vote due to their self-perception as part of an influential group. The decline of each of these potential causes at the individual level led to a collapse in voter turnout in the aggregate, despite the increase in education, which should lead to increased participation (Wolfinger and Rosenstone 1980), which was considered alongside the controls that Wolfinger and Rosenstone and others suggested as relevant. This allowed for the consideration of the hypotheses of several different approaches, but did so at a cost to the overall work's parsimony and interpretability.

Despite Teixeira's consideration of a wealth of causes, it's not clear that each potential theoretically driven causal mechanism is given a fair hearing. Race, region and sex were all dismissed, included as controls despite the author's skepticism about their relevance and the literature's theoretical explanations for their inclusion. This is reasonable, however, given the later findings of Verba et al. who suggested that race did not matter in questions of turnout (1993). In this case the work fell victim to a requirement of the literature, whose discussion of an excess of causes led to a set of models with an excess of variables. But what else could be done? The specification of such models and findings of significance for these variables indicated a need for additional predictors. Even as the models grew to sizes that made their findings difficult to parse, the list of theoretical causes continued to grow, decreasing the likelihood of paring down the theory to a manageable set of explanations.

## Mobilization Theory

In addition to the challenges of the prior literature, and the failure to trim down the list of previous causes, the addition of further causal mechanisms expanded the literature, branching out without cutting down on the number of causes. Such an addition was made by Rosenstone and Hansen, who suggested a new causal mechanism to explain turnout. While an individual's political efficacy and resources are important to the mobilization model, Rosenstone and Hansen's contribution was the inclusion of political actors who could influence the individual. An individual can participate of their own volition, but if the individual lacks the resources necessary to participate, a political organization may step in to bear individual costs of participation. The act of direct contact by one of these organizations lowers the costs of participation for a citizen to the point where they may be able to participate. Rosenstone and Hansen concentrate on political parties for much of their work, but concede that parties are one of many potential mobilizers. In the mobilization model, social organizations such as civil rights groups and unions are also capable of mobilizing voters (2003), allowing other organizations to drive an individual to participate in politics.

In addition to the direct mobilization that political parties provide, Rosenstone and Hansen suggest that citizens are also mobilized indirectly. Drawing partially on the work of Olsen (1971), Rosenstone and Hansen posit that politically active friends and neighbors may decide to treat voting as a collective action problem. A voter embedded in a network with such individuals receives selective rewards for participation, as well as selective penalties for failure to participate in the electoral process. For this reason, individuals embedded heavily in social

networks, indicated by mechanisms such as employment, social group participation, and class, are more likely to be mobilized to participate (Rosenstone and Hansen 2003). For the mobilization model, the act of voting is driven not only by resources and efficacy, but by the mobilization created by political organizations and other actors.

Apart from the concerns of demographically oriented studies, Rosenstone and Hansen contribute a new causal mechanism, avoiding the pitfalls of the literature that concentrated on an increasingly large subset of demographic groups. Despite this necessary contribution, however, mobilization theory leads to a series of other problems. While direct mobilization may be easily captured by questioning people about political contact, indirect mobilization remains challenging to operationalize. The authors argue for social group activity as an indicator, but both mobilization and social group participation may be driven by separate causes. Perhaps the personality trait that leads people to be outgoing may serve as a driver of political activity, or social groups are a natural target for direct mobilization by political groups, making the individual a subject of direct mobilization regardless. While the potential problems of operationalizing these aspects of mobilization cast doubt on the theory, the main concern remains the addition of extra causal mechanisms, as opposed to negating the large body of causes that already exist.

### Additive Contributions and Habitual Voting

Despite the near-consistent attempt to add, rather than reduce the number of predictors, there has been some effort to pare down the number of theoretical causes of voting, such as the

research of Verba et al. (1993). The habit of the literature, however, appears to be the expansion of theoretical causes. Rather than attempting to pare down theoretically specified causal mechanisms, creating a small but highly relevant set of predictors, researchers focus on adding causes. The primary goal of the turnout literature concentrates on finding alternative explanations for turnout, even as this adds to the list of possible reasons a 'true model' would require to explain why someone chooses to vote. This trend can be seen in an offshoot of the psychological literature of the choice to vote, which suggests that the turnout is habitual (Brody and Sniderman 1977, Gerber, Green, and Shachar 2003). In some cases, this leads to sidestepping the theoretical underpinnings of turnout and habit entirely, even mathematically simulating multiple psychological driving mechanisms behind the choice to vote (Fowler 2006). It appears to be adding to the literature, but the contribution of habitual turnout eschews theoretical causes in favor of unnecessary methodological novelties.

While “voting is for many a habit”, as Brody and Sniderman describe it (1977, 349), the literature demonstrates a struggle to explain the underlying reason that habit affects an individual's choice to vote. When researchers consider why voting might be habitual, however, the theoretical drivers of the choice to vote fail to include convincing explanation. In the case of Gerber, Green, and Shachar, individuals have a psychological impetus to vote over time, one which is verified experimentally (2003). While tracing the idea that voting is a habitual behavior to Aristotle, however, Gerber, Green, and Shachar never isolate the reasoning behind habitual voters, instead identifying a wealth of causes with a verifiable effect.

The habitual choice to vote is experimentally verified by Gerber, Green, and Shachar, but

their explanation of why voting is habitual is lacking, due to their ambiguous stance on the theoretical reasoning behind their findings. Voting may be habitual due to individuals' enduring response tendencies; it may be that the same stimuli cause the same result again and again, due to the fact that voters exist in a persistent electoral environment; or it may be due to a self-reinforcing effect, where the choice to vote in one instance leads to an increased likelihood of voting in the future. Of these three explanations, the first two appear to be akin to one another. Individuals vote due to their enduring response tendencies, something which is consistently stimulated within a persistent electoral environment. Meanwhile, the third explanation is a restatement of the original idea, that individuals vote because voting is habit forming. The idea of voting as habitual may be buoyed by experimental evidence, making it difficult to argue that voting is habit forming, but at the same time, the explanation for why voting is habit forming is lacking, a difficulty that makes considering the habitual nature of voting challenging.

Despite the difficulty in parsing the causal mechanisms behind the idea of voting as a habit, there is precedent in demonstrating a new method on a question related to the choice to vote. Fowler, for instance, demonstrates a series of techniques through simulation designed as a formal model of the choice to vote (2006). Beginning with a critique of the simulation effort by Bendor, Diermeier, and Ting (2003), Fowler suggests a correction to their behavioral model that better captures the habitual nature of voting. By adjusting feedback of voting, Fowler creates a model that captures the possibility that an individual may not vote, while also capturing a majority of simulated voters' behavior.

By correcting the feedback simulation of Bendor, Diermeier, and Ting, Fowler appears to

capture the behavior of voters as it appears outside the simulation. It may be, however, that Fowler is talking past Bendor, Diermeier and Ting, due to the possibility of different ways of considering voter turnout, especially if different types of elections are factored in. To use an example, an individual may vote in a specific subset of all elections such as elections for president, satisfying the criterion that Fowler uses as per the theoretical concept of voting as a habit. For this reason, it may be that the two researchers each have much to contribute, but if different conceptions of turnout are differentiated.

When considering how Fowler, in comparison to Bendor, Diermeier and Ting, simulate the choice to vote, it may be that Fowler created the more realistic simulation. Alternatively, Bendor, Diermeier and Ting's simulation may better reflect the large number of elections that voters could participate in, and attempt to simulate a voter's turnout, perhaps in Presidential elections but eschewing smaller electoral contests such as local or special elections. The two differing conceptions of the choice to vote are based purely in simulation, making it difficult to determine whether such different considerations of vote choice are more accurate for each simulation. Although it is possible that only one of these approaches represents a more accurate way of modeling vote choice, multiple types of turnout may mean that for differing choices on voting, the simulations are appropriate at different intervals.

Regardless of whether voting is a habit, the main concern of including prior turnout in models of the choice to vote is its complete overlap with other potential predictors. An individual with increased education is more likely to turn out, as is a person who has a position whose class makes it easier for them to vote (Wolfinger and Rosenstone 1980, Teixeira 1987). The problem

lies in determining whether the person's status as a voter last time differs from their status as an educated, wealthy person, affecting their likelihood of turning out in the same manner. The difficulty then, of determining whether voting is a habit from this data lies in the difficulty of disentangling these different ways of predicting whether a person votes or not. In addition, the theoretical underpinnings for habitual voting do not appear to be strongly specified, especially in comparison to the mobilization, psychological, and sociological theories of voting. For this reason, it is a good idea to focus on fully theoretically specified variables as a means of determining whether their explanations are sufficient to predict, rather than explain the choice to vote, putting habitual turnout aside until a better theoretical explanation can be provided.

## A Literature That Adds, But Does Not Weigh Explanations

In the consideration of each of these differing conceptions of the choice to vote, the problem of the literature is its inability to parse the different theories, to determine which of the conceptions of turnout best reflects what is, rather than what is theorized. This is not reflective of the work of theorists, whose contributions are invaluable, but the means by which the different theories are tested. One of the early critiques made of the social sciences in general, is why multiple phenomena cannot contribute to a topic of interest in equal measure. While the equivocation of different causal mechanisms is not particularly useful due to the fact that it adds little to the understanding of a research topic, the problem of figuring out which theoretical approach provides better predictors of a phenomenon remains unanswered.

While the literature is capable of developing endless theoretical explanations, the chance

to force them to compete with one another remains untaken. Too often, the literature focuses on adding, rather than subtracting elements of the models, until giant lists of variables are necessary to suitably control for the collection of theoretically suggested causes, akin to the diverse areas studied in interest groups, with little true accumulation (Baumgartner and Leech 1998). Such an approach to determining veracity falls decisively in opposition to parsimony exemplified in Achen's rule of three. To Achen, any model with more than three independent variables is meaningless due to poor specification, arguing that treating the groups within enormous models as something that can be controlled for through dummy variables is incorrect (2002). But the main thrust of the literature does not appear to be to answer Achen, either by arguing against his rule, or by trying to work within it, but to ignore his critiques of research methodology entirely (Schrodt 2014). For this reason, a different approach is needed to

While a few efforts, such as Verba et al. (1993) have been made to cut down on the number of variables that explain political participation, the main thrust of the literature is to continue expanding lists of relevant variables. The literature focuses on building, rather than sculpting the collection of variables that should be included in understanding the turnout. While multiple causes may drive the choice to vote, and may even simultaneously explain parts of the variation in turnout, a predictive approach such as BeSiVa, similar to other variable selection approaches, allows for a comparison of variables equivalent to sculpting, rather than accumulating. BeSiVa compares the variables which are provided to it, using the information to create a model that best predicts subsets of data held out for the purpose of prediction.

## The Data

In order to understand what drove voter turnout, the American National Election Study, or ANES, was selected as a source for data. The ANES serves as a logical choice to study turnout, especially at the level of individual voters. The study's time series cumulative data file allows for consideration of differing drivers behind the choice to vote ranging back to 1948. In addition, the data allows for a consideration of a large collection of potential drivers of turnout in a manner that lets BeSiVa make a comparison of each predictor's relevance. Once any category deemed missing was recoded appropriately, it was necessary to consider what data should be chosen to use with the algorithm. While BeSiVa could work on any data set, a selection was needed, and the 2000 election was selected at random from a set of possible years. Like the other possible datasets, the ANES 2000 survey contained the variables suggested by theory, had a sizable number of observations, and was readily available.

## The Dependent Variable

Given the focus on the choice to vote, the operationalization of the dependent variable was relatively straightforward. In one of their survey responses, the ANES asked whether an individual voted in a given election, which was used as a proxy for their behavior. There is an understanding, however, that an individual might decide not to vote, but when queried, would specify that they did due to response bias, (Belli, Traugott, and Beckmann 2001, Tourangeau and Yan 2007). Despite this potential drawback of using survey responses rather than voting records, the risk of response bias is accepted as a limitation of the design, and is irrelevant to the

algorithm's ability to predict. Given that the main point of using turnout is to test the algorithm, the risk of response bias in the dependent variable is secondary to the utility of the predictions created. Missing data were properly recoded, and the survey responses were changed into a numeric dichotomous variable with 1 for yes and 0 for no, but no other changes were made to the dependent variable. Having selected turnout as a dependent variable, the algorithm only needed a series of independent variables to create a predictive model of the choice to vote.

### Independent Variables

Having determined that self-reported choice to vote would serve as the dependent variable, multiple independent variables were provided to BeSiVa for its consideration. The first variable that was included for consideration was an individual's party identification. Dating back to the Michigan School, party identification may affect an individual's likelihood to turn out due to interest and attachment to the outcome (Campbell et al. 1960, Dalton and Wattenberg 1993). In the ANES data, party identification was measured as a 7 item categorical variable, ranging from strong partisan, weak partisan, independent but leaning towards a party, and truly independent. For the algorithm's consideration, and due to the fact that party strength, rather than party, is theorized to lead a person to turn out to vote, however, party identification was recoded as a numeric variable capturing the strength of an individual's identification. This operationalization of the strength of party ID ranged from 0 for independents to 3 for strong partisans of either party. Thus, party identification, long associated with turnout, was included in a manner consistent with the component expected to correspond to turnout behavior, along with other variables suggested by the literature.

In addition to party identification, a collection of variables based off the opinions of sociological theorists were included, starting with education. Operationalized based on if someone had achieved certain levels of schooling, Wolfinger and Rosenstone theorize that education increases how much individuals pay attention to politics. Education makes politics more enjoyable to follow, as further additional education makes it easier to understand the rules and impact of politics (1980). It has also been suggested that education aids individuals in maneuvering the mechanics of voting (Teixeira 1987), or enables the individual to better parse political information (Rosenstone and Hansen [1993] 2003). Education was included in the ANES as a 6 category factor variable, ranging from finishing grade school to advanced degrees, with values in between to capture whether someone had finished high school or a bachelor's degree. The variable, however, was taken and transformed into a numeric variable, one which captured a person's level of education as a continuous predictor. This operationalization of education was relatively straightforward, and its multiple backing theories and relative agreement made it a necessary addition to the list of variables to consider.

Education was one example of a variable with multiple proposed causal mechanisms related to turnout behavior, as was age. Age was hypothesized to affect voter turnout as a proxy for political experience, which could substitute for education (Wolfinger and Rosenstone 1980). It was also a potential driver due to the existence of shared generational experience or differing drives at stages of the life cycle (Rosenstone and Hansen [1993] 2003). Age was asked about on the ANES as a numeric variable, and was provided to the algorithm in its unrecoded form, as well as an age squared term, to capture possible non-linearities. This meant that age was included based on how old the individual was at the time of the survey, a continuous predictor of an

individual's experience.

In addition to variables that indicated a person's cognitive resources, such as age and education, the literature described individuals' connections to media as a possible determinant of turnout behavior. Teixeira suggested that links to the media, operationalized based on how often someone read the newspaper, allowed voters to be more engaged, giving each election a sense of meaning and making an individual more likely to turn out (1987). The ANES inquired about newspaper reading directly, and the number of days per week an individual read the paper was provided to the algorithm as a possible predictor of turnout. The number of days reading the paper was included as a numeric predictor, and ranged from none to seven. With this way of determining how politically connected through the media an individual was, the algorithm could capture these links' overall relevance.

Due to the fact that it has been defined in multiple ways, political efficacy was challenging to include for the algorithm's consideration. Political efficacy, based in the notion that an individual believes they influence the outcome, has been theorized to increase the likelihood of voting due to the increased sense of accomplishment that such a feeling provides (Teixeira 1987). It can also make things easier for a potential voter, with a sense of personal competence makes someone feel more comfortable participating in politics overall (Rosenstone and Hansen [1993] 2003). Political efficacy is operationalized using this sense of influence over the outcome, helping to determine its role in the prediction of whether a person votes by including it for the algorithm's consideration.

Despite the fact that it has been suggested as a key driver in some corners of the literature, the role of income in predicting turnout has been disputed. While such a resource may make individuals more capable of participating in politics (Schattschneider 1975), Wolfinger and Rosenstone suggested that income only mattered to the point of comfort (1980). Teixeira, by comparison, suggested that greater incomes were capable of easing the challenge of voting as a component of socioeconomic status (1987), while Rosenstone and Hansen suggested that it would make participation more likely, due to an increased likelihood of sharing social circles with the political class ([1993] 2003). To try and capture income's potential role, the income categories of the ANES were recoded as a numeric predictor, treating the different selections of the respondent's income quantile as a driver of political participation. Once it had been recoded, its ability to capture whether income could potentially predict turnout made it an invaluable addition to the list of variables for consideration.

The use of race as a predictor of whether someone turns out to vote has been disputed (Verba et al. 1993) as are individuals' sex and the region in which someone lives (Teixeira 1987). Their roles as operationalizations of historical privilege and political culture (Wolfinger and Rosenstone 1980, Teixeira 1987), however, and their disputed status meant that these demographic predictors should be included in the list of variables to consider. Race was recoded from a 7 point scale based on a self-identification of ethnicity, which was changed to a dichotomous variable based on whether an individual was in the minority. Region was a 4-fold classification based on whether someone lived in the northeast, north central, south, or west of the United States. Similarly, sex was included for consideration as a binary dichotomous categorical variable. Given their disputed status, these variables are ideal for a predictive

algorithm, allowing for their consideration in a more systemic fashion.

Unlike the disputed status of race and other demographic predictors, marital status is a relatively uncontroversial addition to the list of independent variables for the algorithm to consider. Being married lessens the costs on an individual to go out and vote while providing a separate incentive to do so due to an additional stake in the election (Teixeira 1987). While not directly mentioned, marital status makes intuitive sense from a mobilization perspective as well. A second person increases the likelihood of being embedded in social networks, increasing the likelihood of direct and indirect mobilization (Rosenstone and Hansen [1993] 2003). To capture marital status, its effect on networks, and where it put someone, an individual's marital history was considered. Although it was asked about in the ANES from the perspective of whether someone had been married, the variable was dichotomized to focus on whether or not someone was divorced, as a way of separating out those who had been married and separated from those who had not. With this operationalization of an individual's overall marital status, BeSiVa could determine its relationship to turnout due to the inclusion of a variable operationalizing it.

Due to their consideration as potential operationalizations of mobilization, a collection of variables was included to determine whether individuals voted due to political contact. Multiple variables to determine whether an individual had been contacted by a party or some other organization were included for the algorithm's consideration, as well as their employment status, a key predictor of being at the center of a network of potential mobilizing agents (Rosenstone and Hansen [1993] 2003). The ANES captured these responses through a series of dichotomous questions, asking whether an individual had been contacted by the Democratic or Republican

parties, any party, or by another political organization. These variables were included for the algorithm's consideration as a means of capturing whether someone had been directly mobilized.

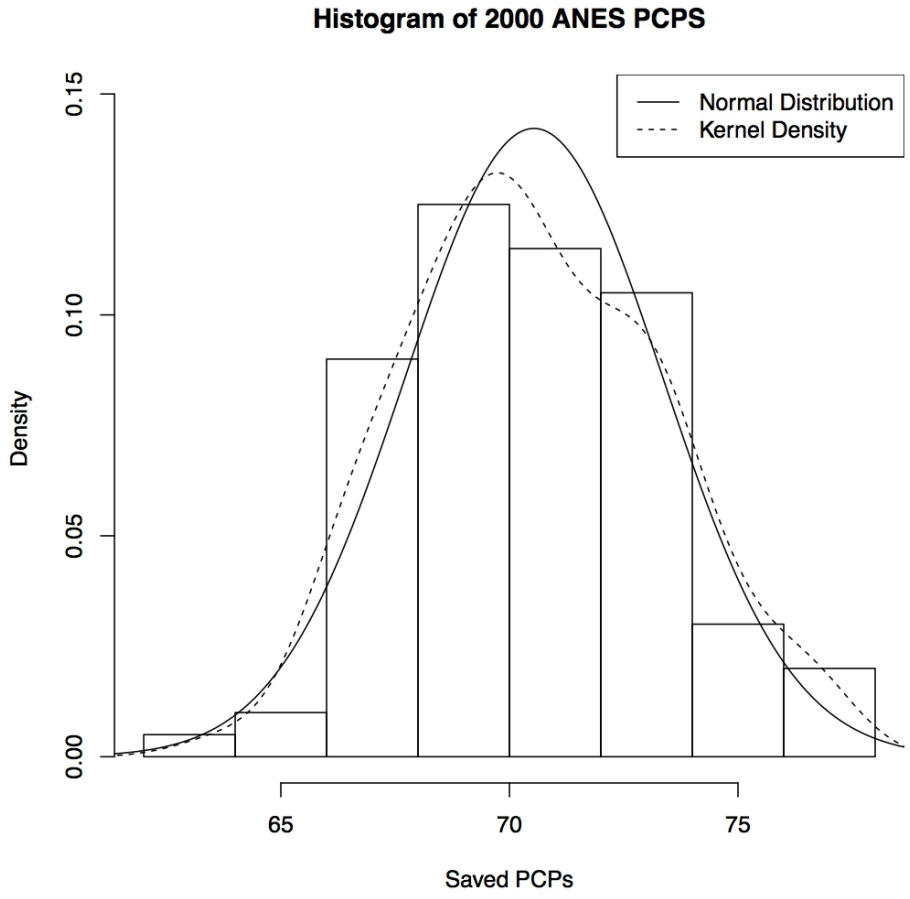
Direct mobilization is a key means of determining participation in the mobilization model of political participation, but it remains half of the main drivers of mobilization theory, and merited inclusion. This was done in the form of an individual's connection to networks through religious participation, asking whether they attended church or not. Church attendance was used to further capture the potential role of such a network, which was asked as a 5-item question based on how often someone went to church over the course of a year. This was recoded, however, into a dichotomous response to determine whether someone attended church or not. With this means of capturing whether an individual was connected to a greater network, a selection of variables most relevant to understanding whether an individual chose to vote or not could be considered by the algorithm.

## Methods and Results

Once the variables were recoded into more theoretically appropriate forms, they were provided to the algorithm, which went to work determining which independent variables best predicted turnout behavior. Instead of running the algorithm once, however, the algorithm was run on the same data 100 times, randomly separating the data into different training and test sets each time. The training and test sets were varied by the use of different random seeds, guaranteeing a different division of data each time while also making it possible to recreate results as needed. The data, independent variables, and dependent variables were always the

same, as were the arguments provided to the algorithm. These arguments specified five iterations, meaning that the maximum number of variables in any single model the algorithm created was 5. It also specified a threshold of 0.1%, meaning that any variable that could be added to the model would need to make the PCP improve by at least one tenth of one percent, or the algorithm would stop and return its results. The algorithm was run on the data, and the results of the PCP as well as the independent variables were saved, analyzed, and compared against theoretically specified models.

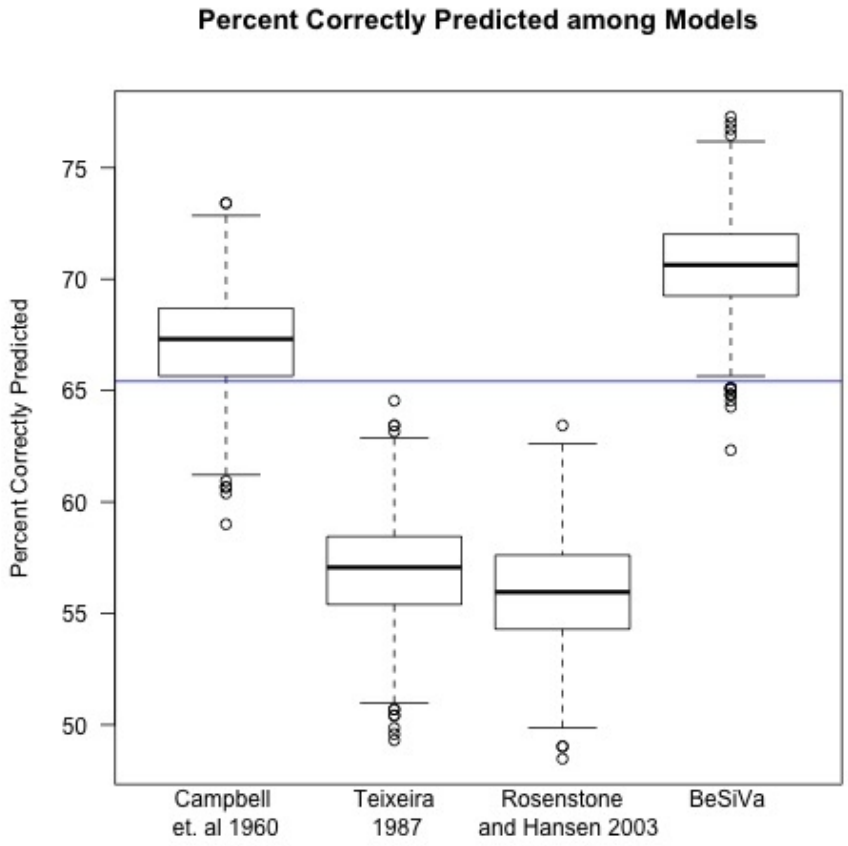
Figure 2: The algorithm's percent correctly predicted. The final models for BeSiVa had a mean of slightly below 70%, and bear superficial similarities to a normal distribution.




From the 100 runs of the algorithm, two of the major outputs related to the final model that BeSiVa recommended, its PCP and the independent variables that it specified, were saved. The PCPs for each of the runs may be seen in figure 1, displayed in the histogram and kernel density plot. This shows the distribution of the PCPs, and notably the mean PCP. On average, the percent correctly predicted of these models is 70.5%, and the median is 70.1%, showing that the models created by BeSiVa are better than a random guess (which would lead to a PCP of 50%),

and better than predicting the modal category of the dependent variable for all voters (which would have a PCP of 65.4%). The predictions are concentrated tightly around the mean, with a standard deviation of 2.8%, suggesting that the predictions the algorithm makes are consistent. In addition, a normal distribution created using the mean and standard deviation of the PCPs has been superimposed over the plot, displaying a marked similarity to the PCPs' kernel density[1]. This plot provides a demonstration of the PCPs from the models that the algorithm generates, which may be compared to the predictions made by theoretically specified models.

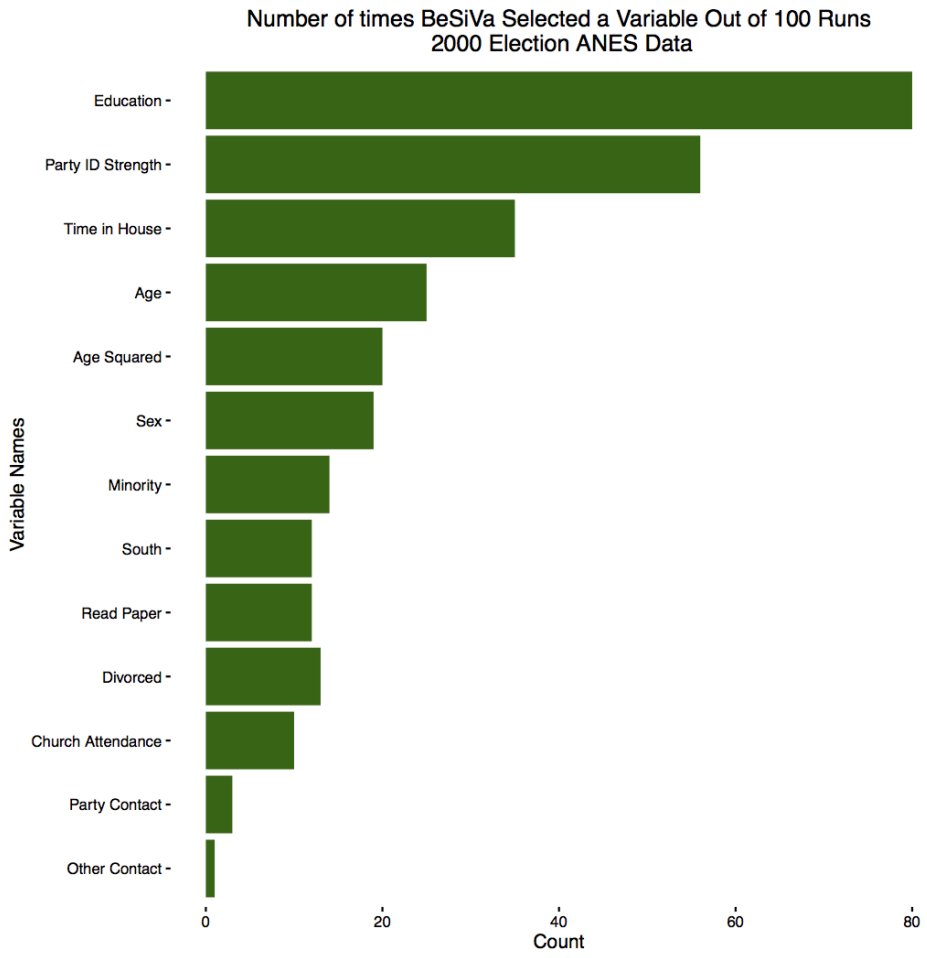
Figure 2: A comparison of the PCPs between theoretically specified models and BeSiVa. The medians are signified by the thick black bars at the center of the boxes. The medians of the theoretical models here fall well below 70%, suggesting that these models are less capable of predicting whether a person votes, even Campbell et al. 1960, whose model is still well below the mean percentage of the percent correctly predicted.




To compare BeSiVa's models to the theoretically specified models from the literature, it is necessary to compare predictions. To make this comparison, the models using variables specified by Campbell et al., Teixeira, and Rosenstone and Hansen were created with 20% of observations held out at random, placing them in a test set. Campbell et al.'s model only included party identification, while Rosenstone and Hansen included efficacy, education, income, age, and the network placement indicators, including church attendance and contact

from parties and other organizations. Teixeira, meanwhile, included a vast number of variables. These variables included the demographic data seen in other models, links to the media, operationalized by the days someone read the paper, political efficacy, and the strength of partisan identification. Just like the BeSiVa algorithm, PCPs were generated on the held out data for the theoretical models. This process was repeated 100 times, and the PCPs for these models and one created from the four variables suggested by BeSiVa were saved and displayed in figure 2. Figure 2 shows a set of box plots which are compared to choosing the mode for all observations. As signified by the blue line, choosing the mode of the dependent variable for all observations would lead to a prediction that is 65.4% accurate, the proportion of people who said they voted. This shows that where accuracy is concerned, only BeSiVa's and Campbell et al.'s variable choices do better than picking the modal category for all observations. The black bars in the boxes represent the median, the central point of the data, allowing the theoretical models to be compared to one created by the algorithm, whose models had a median PCP of 70.1%. The box plots show that the median PCPs generated by the theoretically specified models are smaller than the median PCP of BeSiVa. This suggests that despite their theoretical specification, these models are not as capable of making predictions as one specified by the algorithm.

Figure 4: BeSiVa's selections. Out of 100 runs, BeSiVa selected education most often, followed by party ID and time spent in someone's house.




In addition to the PCPs, the algorithm also provided a selection of independent variables in its final model, those that maximized the percent correctly predicted. These variables were saved and plotted in figure 3, with the x axis representing the number of times each variable was selected, and the y axis naming each independent variable selected by the algorithm. Unsurprisingly, the most commonly selected variable for predicting whether an individual turned out to vote is their level of education, chosen in approximately four fifths of the algorithm's

selections of independent variables. Following closely is the strength of an individual's party identification, chosen over half of the time. These two variables are chosen often enough that based on the strength of the theories underlying their inclusion, and their performance in the algorithm, any model created from the results of BeSiVa ought to include education and party identification.

While education and party identification are the two most commonly included predictors in the algorithm's results, it is hardly surprising that these two well explored, theoretically validated variables were found to be useful in predicting vote choice. Surprisingly, however, the operationalization of mobility, the time spent in one's house, is the third most selected predictor, suggesting that residential mobility is important in determining an individual's likelihood to $vote_2$[2]. The time spent in one's house is highly favored by the algorithm as a predictor, selected third most often of all included independent variables. Perhaps this is due the entrenchment of individuals in a network, making them more susceptible to pressures to vote (Rosenstone and Hansen [1993] 2003), or due to more prosaic concerns. After all, moving requires that a person change their voter registration, a barrier to voting that may lead individuals to stay home on election day. Regardless of the theoretical underpinnings, BeSiVa selected the time an individual has lived in a house as the third most important predictor in vote choice, necessitating further investigation.

After time spent living in the same house, the rest of the independent variables selected by BeSiVa include few surprises, except perhaps for the fact that after party identification, a predictor favored by the psychological approach (Dalton and Wattenberg 1993), BeSiVa has a

definite preference for the sociological approach. The age and age squared terms are selected, demonstrating the predictor's strength in predicting likelihood to vote. The algorithm prefers other demographic predictors, such as an individual's sex, status as a minority, and region, suggesting roles for these sociological predictors in determining vote choice. After the main demographic variables, two of the Mobilization theory's predictors are considered, but these variables were selected in less than 5 of the 100 times the algorithm was run, suggesting that their selection is unlikely to predict turnout in the majority of cases. This test demonstrated the relative quality of the variables selected, suggesting that among those tested, the psychological and sociological theories of vote choice were most relevant for predicting turnout.

### Model Validation Through Bootstrapping

Having obtained a sense of where theories stood for predicting whether someone chose to turn out or not, the question of how to determine the quality of the variables and models selected by BeSiVa, and compare them in a systemic way to the theoretically specified models became imperative. After all, BeSiVa had provided 100 separate recommendations for independent variables, each of which included a slightly different selection of variables provided to it. In an attempt to establish the quality of models according to the predictive criterion, a process similar to BeSiVa was attempted, comparing models with the variables it suggested against models containing theoretically relevant predictors, which can be seen in figures 4-5. The results suggested the remarkable power of education in predicting whether or not someone chose to vote, and the difficulty in determining the necessity for predictors beyond 4.

Figure 5: A test of the independent variables. The independent variables selected by BeSiVa in figure 3 were added to models one at a time, and cross-validated 100 times. In this case, the best model is the one that has 7 variables included, but the variation is large.

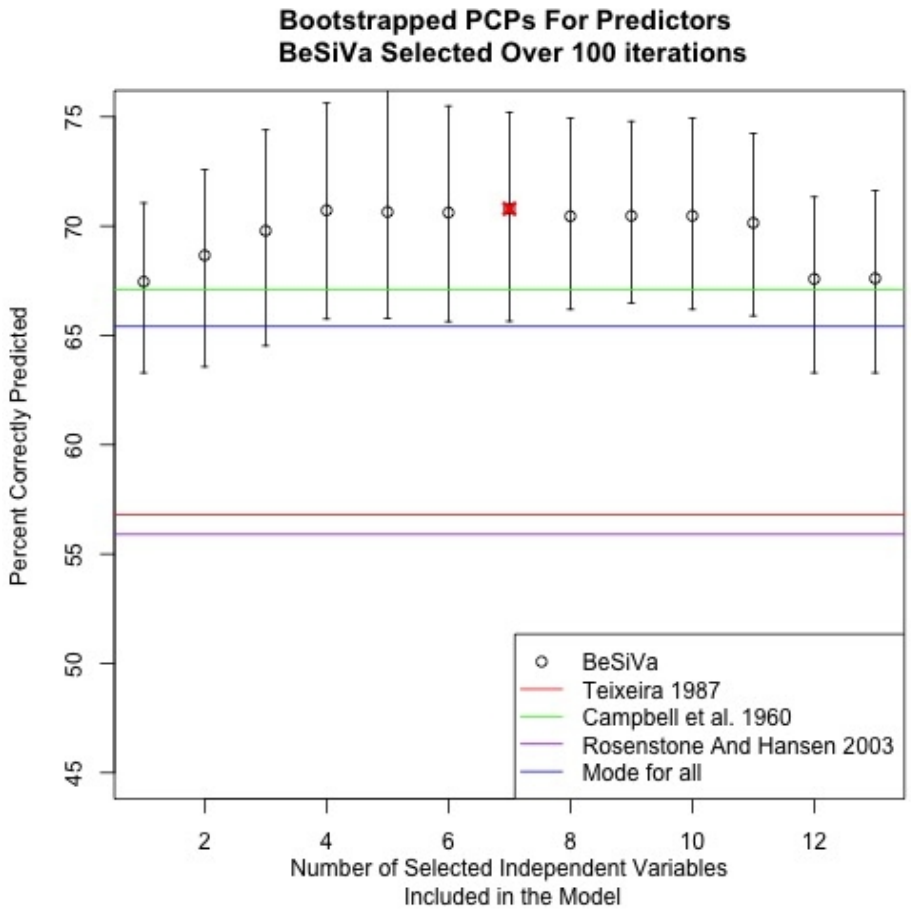
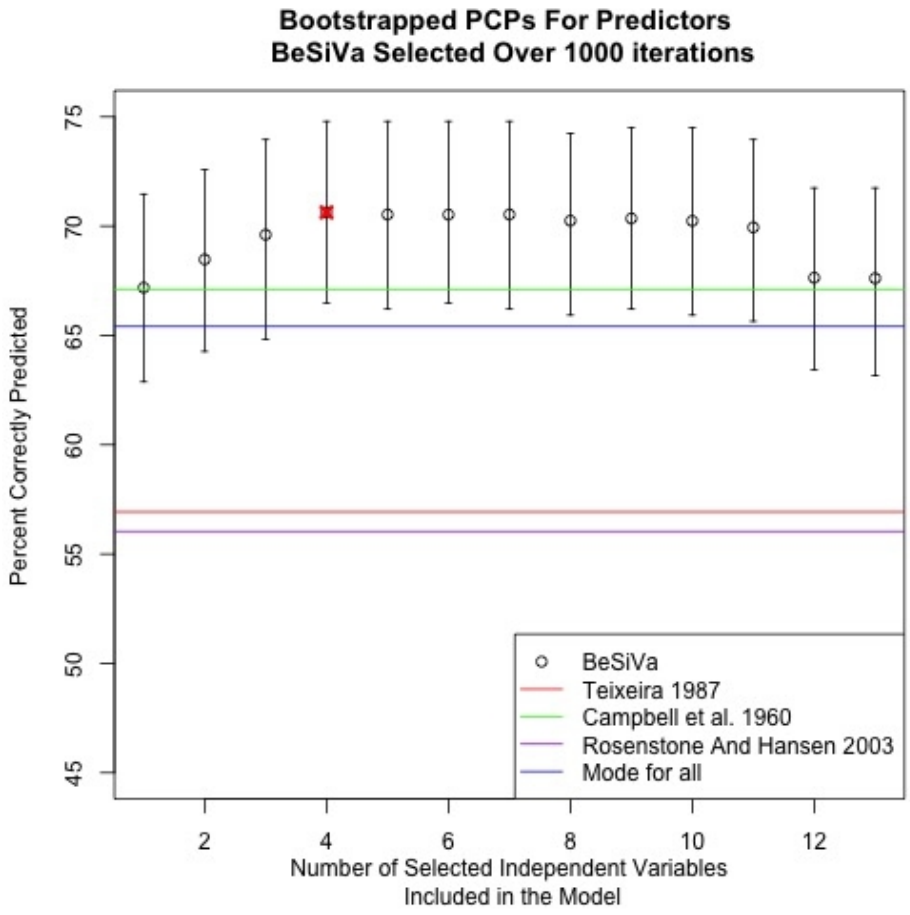


Figure 6: the same test, repeated 1000 times. Note that in this case, the largest PCP is also the first one whose 95% confidence band fall outside the choosing the mode for category, with the first 4 independent variables.




To determine the suitability of the independent variables selected by the BeSiVa algorithm, the predictors were listed in the order of the number of times they were selected, and then added to models. The first model contained education, while the second contained education and party identification, following the pattern of variables seen in figure 3. Each of these models was subject to a cross-validation strategy, similar to the process of BeSiVa. A random subset of 20% of observations were kept from each model, and the model was reestimated without that

data a specific number of times. Then, the held out data was predicted and compared against its measured value. This was done for models featuring each of the independent variables or variables selected by BeSiVa, and for three models based off of theory drawn from the mobilization, sociological, and psychological perspectives of turnout.

Once the models were tested and compared, their average PCP were stored, and plotted as the points in figures 4-5[3]. There are a few things to note about these figures. First, there is the relative similarity of all of the models predicted by BeSiVa. Despite adding upwards of 13 variables, the models all have a relatively similar range of PCPs. The model with the largest PCP, signified by the X, appears to decrease from 7 to 4, as seen in the figures. While this might indicate that as the number of bootstrapped runs increases, the large confidence bands suggest that the maximum PCP is barely decreasing. Despite this, a confident statement may be made about the number of variables necessary to make a good prediction based on these confidence bands, due to their increase above the prediction that would be made if the mode were chosen for all observations.

If the mode were chosen for all observations, as displayed by the blue line, the percent of correctly predicted observations would not vary. Specifically, it would fall at 0.654, the proportion of individuals who said they voted in the ANES' 2000 Survey. By picking this value for all voters, however, someone attempting to make a prediction would outperform theoretically specified approaches, such as those suggested by Teixeira and Rosenstone and Hansen, as seen in the lines near the bottom of the chart. The addition of many variables, as Teixeira (1987) felt required to do, made for poor predictions, and the model suggested by Rosenstone and Hansen's

mobilization approach made a similar prediction. The Michigan school's contribution, that the strength of party identification makes an individual more likely to turn out, nearly outperforms the initial model, but is quickly outperformed by BeSiVa. With the exception of the Michigan model, the highly theoretically specified models are incapable of making better predictions than choosing the modal category for all voters.

Although picking the modal category for all voters outperforms two of the theoretically specified approaches, it also may outperform some of the models created by BeSiVa. Given that in figures 4-5, the mode occasionally falls within the 95% confidence intervals, the models created with variables suggested by BeSiVa may not always outperform choosing the mode in all cases. Despite this possibility, however, some of the models do outperform choosing the mode. Figures 4-5 all show that after 4 variables are added, the confidence band falls above choosing the mode, which continues until ten to twelve variables are added to the model, depending on the number of times cross-validation was run. In the interest of parsimony, it may then be suggested that four variables, education, party identification, time spent in one's house, and age, theoretically specified and chosen by the algorithm, maximize prediction on the choice to vote using BeSiVa.

## A Comparison to Statistical Significance

Having demonstrated the results suggested by the algorithm, it makes sense to compare them to the results generated by null hypothesis significance testing, examining the statistical significance of the variables and how the theoretically specified models differ from those the

algorithm proposes. Instead of running bootstrapped regressions to compare predictions, each of the models suggested in the last step were run as separate logistic regressions, as were regressions featuring the variables recommended by Campbell et al. (1960), Teixeira (1987), and Rosenstone and Hansen (1993). All available data were used for these regressions, with no separation of data for testing purposes, and the models resulting from each of these regressions may be seen in tables 1-3. Comparing the models of these tables, it is clear that statistical significance, while capable of explaining variation in the dependent variable, makes it impossible to determine a variable's ability to make predictions.

The models seen in tables 1-3 have many variables with statistical significance, but the variables that significance suggests are relevant frequently make the predictions worse. Although many of the models that seem most appropriate for predictive purposes have significant variables, not all of the variables recommended by the algorithm for making predictions are statistically significant. The first four most commonly selected variables all are statistically significant, suggesting an accord between the algorithm, bootstrapping, and significance testing. The remainder of the models, however, display scattered statistical significance, some with p-values well under the widely accepted threshold of 0.05, despite the fact that no model performs better than the one created by the first four recommended variables according to the 1000 run bootstrap or seven variables according to the 100 run bootstrap. In addition to the quality of fit statistics, Tables 1-3 feature the average PCPs from the 1000 run bootstrap, which are also included separately in table 4. Regardless of which model is capable of making the best predictions, the scattershot nature of statistical significance makes it difficult to determine which variables are truly useful in predicting turnout. With some variables, their inclusion appears to

make sense from an explanatory perspective, but when their effect on the average PCP is considered, it is clear that these variables' inclusion make a model's ability to predict worse.

Depending on the number of bootstrapping runs to generate the confidence intervals on the PCPs, an argument may be made about whether the seventh or the fourth model is more capable of making predictions. Looking at statistical significance, however, is counterproductive if the goal is to make a prediction from the model. As an example, the role of the squared term in age is debatable; in the 100 run bootstrap, it appears to lead to a maximizing of prediction, but the small coefficient size and varying significance make it difficult to determine whether it should be included. An individual's minority status appears to affect turnout, and it is significant until party contact is added, but this suggests that contact supersedes an individual's racial identity in a model that predicts whether a person turns out. This is especially troubling, due to the fact that based on the bootstrapping tests and the mean PCPs generated from those tests, party contact's inclusion leads to empirically worse predictions. Similarly, gender is never found to be statistically significant, despite improving models' predictive power according to the 100 run bootstrap. It may be that the 100 run bootstrap is doing a poorer job of determining which model is most capable of predicting turnout, but statistical significance decisively misleads the researcher, overstating the importance of predictors like the days an individual read the paper, church attendance, and especially party contact. The use of statistical significance may inform the researcher if a non-zero relationship exists, but the idea that p-values inform the utility, substantive significance, or even relevance of a particular independent variable on predicting the dependent variable is hopefully abated.

## Quality of Fit Statistics

Although significance demonstrably misleads if a prediction is to be made, this is not necessarily a problem with significance itself; just because a variable is considered significant doesn't necessarily mean it is supposed to be predictive. This is contrary, however, to the way that statistical significance is frequently considered. Too often, NHST and quality of-fit statistics and predictive capability are conflated (Shmueli 2010), and the quality-of-fit statistics confuse the issue further. Looking at models' quality-of-fit statistics, it is clear that despite the algorithm's prescription for parsimonious models, the estimates of the models' deviance and Akaike Information Criterion (AIC) do not change in a manner that reflect the predictive capability of each model. If used to estimate predictive capability, the quality-of-fit statistics suggest that the models are improving at making predictions, decreasing consistently (with a slight exception) as variables are added to the models. Given that these statistics are meant to suggest the difference between models' performances, serving as measures of goodness of fit (Long 1997), this is especially problematic given that they are thought to signify a model's ability to make good predictions (Shmueli 2010), which is explicitly stated for the Akaike Information Criterion (Forster and Sober 1994, Forster 2002). Given the utility of the models suggested by the algorithm, and the predictions made by the models in the bootstrapping, a model's ability to predict and its ability to explain, as signified by the quality-of-fit statistics, must be explicitly differentiated.

As previously described, tables 1 and 2 feature a set of regressions based on variables suggested by the algorithm, and they include the quality of fit statistics for these models, as well

as their mean PCP. Now, if the 100 run bootstrap is considered, then the model with seven variables is the first to outperform choosing the mode for all voters. If the 1000 iteration bootstrapping process is considered, then the four variable model features the only four variables necessary to create the best prediction of the choice to vote. This would be difficult to determine from looking at the quality-of-fit statistics, however. These estimates of model quality are supposed to measure how well the model explains the provided data, or in the case of the AIC, how well the model separates the predictive trend from the noise in the data (Forster and Sober 1994, Forster 2002), an explicit measure designed to predict new data comparable to the PCP (Clark 2004). Their performance, however, indicates the need for a prediction as a separate criterion based on how they improve, nearly consistently, given additional independent variables.

In considering the quality-of-fit statistics for each model, it would be impossible to determine that the best prediction, on average, is made by a model with four to seven independent variables when looking at the deviance. Despite the fourth model's status as the most predictive model in the 1000 iteration bootstrap, the deviance of each model continues to decrease even after variables are added that do not improve a model's predictive capability. Deviance, a comparison between a perfect prediction of the results and the model created and displayed, is meant to compare between models, determining which ones are best at fitting the data (Long 1997). In this case, however, models with variables that lead to demonstrably poorer predictions still have lower deviance. Instead of one of the algorithm's suggestions, relying the deviance would suggest that the most predictive model is either one proposed by Teixeira or Rosenstone and Hansen, if it is interpreted as a measure of prediction. The deviance almost continuously drops as variables are added, sometimes dramatically even if the prediction is

worse, suggesting that deviance cannot be used to determine which model makes the best predictions. The deviance may be useful from an explanatory standpoint, but it misleads a researcher who seeks to use a model for making predictions.

In considering the models' summary statistics, deviance is not useful for determining whether a model's predictions are better or worse than other models. The deviance is likely overfitting due to its purpose, which is to compare the relationship between the model at hand to a model that is capable of perfectly predicting the data. The Akaike Information Criterion (AIC), however, is meant to be a measure of model performance for predicting data that was not included (Forster and Sober 1994, Forster 2002), and its use as an information criterion has been suggested in contrast to using a test set for making predictions (Clark 2004). Despite this formulation, however, the AIC experiences a problem similar to the deviance, suggesting that it too is vulnerable to overfitting. In tables 1 and 2, the AIC continues to decrease, suggesting that the predictions that each model makes are improving, and the AIC decreases dramatically when contact by a party is added to the model. By comparison, the average PCP is direct a measure of the predictions that each model makes, and it suggests that once 4 variables are added, no improvement on prediction is possible. Just like the Deviance, the AIC suggests that there is a nearly consistent increase in the quality of predictions as variables are added, with only a slight hiccup when the variable accounting for the south is added. By comparison, the PCP suggests that the reverse is true; adding many of these variables leads to poorer predictions, suggesting that similar to the deviance, the AIC cannot determine a model's predictive capability despite its specification as being able to do so.

When comparing the models suggested by the algorithm, the AIC's value drops when almost every variable is added, despite the fact that adding more variables to models made their predictions worse, as demonstrated by the bootstrapping. The AIC continues to decrease in value among the models built from theoretically specified variables, with the exception of the Michigan school's recommendation to consider strength of party identification. Notably, the lowest value for the AIC among all of the explanatory statistics is held by the model with variables suggested by Teixeira, which had a lower predictive accuracy than any model suggested by the algorithm, and is comparable to a model containing the variables recommended by Rosenstone and Hansen. The PCPs, however, actually decreased, showing that the AIC was underperforming as a way of determining how well a model could make predictions. The AIC also suggested that of the three theoretically specified models, the best model for making predictions was Teixeira 1987, with Rosenstone and Hansen 1993 in second place, and Campbell et al. 1960's predictions performing the worst. Empirically, the reverse is true, with the Michigan school doing a better job of making predictions than the other two models in table 3, as seen by comparing their average PCPs. When any measure, statistical significance or quality-of-fit, is compared to actual predictions, it becomes clear that no tool currently used by political science is capable of making accurate, relevant predictions. This calls for an addition to political science's toolbox, tools that focus directly on making predictions. This chapter demonstrated two such tools, and how they can be used in a deductive mode of inference: BeSiVa and the bootstrapping approach. With these techniques, the question of where relationships exist may be considered, but also what is predictive, and therefore what is useful in determining the value of a given dependent variable.

## Conclusion:

This chapter represents a demonstration of the utility of the predictive approach when employed with a deductive research design. This demonstration involved collecting a series of theoretically relevant variables and providing them to the BeSiVa algorithm. Once the variables were provided, the algorithm used its predictive criterion to determine which variables were necessary to create the most predictive model. From there, the variables it recommended were added to a model, one at a time, testing their utility separately using a bootstrapping approach. In addition to the algorithm's variables, models were created from three different theoretical perspectives of the choice to vote, and they were compared to the algorithm's recommended variables, via bootstrapping and the more conventional regression tables. These tables only served to demonstrate how a model may possess statistically significant variables, with a veritable constellation of stars suggesting significance, while making predictions that barely differ from flipping a coin. Based on the algorithm and the bootstrap, four variables appear to create the most predictive model: Education, the strength of party ID, how long someone lives in a single house, and age.

In addition to the comparison of different theories behind voter turnout, using the algorithm enables a consideration of several methodological critiques of the discipline. The idea that a well specified model contains only three independent variables (Achen 2002) may be accurate, but additional independent variables may add to the model's utility from a predictive standpoint. Whether such an addition is warranted requires testing the models using the algorithm and bootstrapping techniques, but through these techniques, it is possible to determine

how preferable parsimony really is. If it is possible to include fewer variables while maximizing prediction accuracy, then parsimony is a useful goal. If the inclusion of a smaller number of variables fails to create a more predictive model, then parsimony is less desirable.

The BeSiVa algorithm and the bootstrapping used to consider the variables it selected systemically demonstrated that in the case of the choice to vote, a model with more than three independent variables may provide additional benefits from a predictive standpoint. But the bootstrapping, which demonstrated how BeSiVa's selections could be used to create predictive models, issues a potential response to a second charge Achen levies. Achen suggests that methods such as logistic regression are overused, and not the most appropriate way to model data. Through predictive criteria such as the PCP, the determination not only of the most appropriate variables, but also the method's appropriateness used may be considered from a more objective standpoint.

While it may be beneficial to consider the use of alternatives to logistic regression, prediction also lets researchers determine the necessity of finding an alternative. If logistic regression leads to good predictions on data that has been kept away from the regression for testing purposes, then a hunt for a better estimator than the logistic estimator is nonessential. If logistic regression does not make good predictions, however, then the methods used to create the model must be reconsidered. Through the BeSiVa algorithm and bootstrapping, however, such a necessity may be determined. This can be done by comparing the predictions against 1/C, where C is the number of categories that can be predicted, as described by Kuhn and Johnson (2013), or by comparing the predictions against the prediction from choosing the mode for all observations.

The fact that by a predictive criterion, such a comparison is possible with logistic regression shows that the estimator has utility beyond what its critics suggest.

The predictive criterion may be used to make the most capable model, using a prespecified technique and variables that are chosen for their theoretical capability. Using such a criterion may even concur with ironclad theoretical findings. But the power of an approach that explicitly considers the ability of a variable, model, or theory to make predictions arises from the possibility that a well specified theoretical model, or a variable whose relevance seemed unimpeachable may not be capable of aiding in actual prediction. This was demonstrated by comparing theoretically specified models and their ability to predict whether an individual turned out to vote, to the models created by the BeSiVa algorithm, and finding the theoretical models wanting. This criterion allows for the consideration of how well models make predictions, making it possible to determine whether theories, either through groupings of variables or the overall selection of what to give to the algorithm, leads to a good prediction. It also demonstrates that statistical significance is incapable of determining what makes a good prediction. Despite the apparent relevance of a given variable according to statistical significance, only an actual predictive test may determine a variable's utility and justify its inclusion from a substantive perspective.

1. While the histogram bore marked similarities to a normal distribution, the bounded nature of the data makes the conventional means of plotting confidence intervals problematic, due to the central limit theorem's lack of applicability to bounded distributions.

2. While such a conclusion validates a hypothesis mentioned by Teixeira (1987), the variable meant to

operationalize mobility in this case may also captures more about the individual than anticipated. Given the lack of class-based variables in the algorithm's selections, despite the inclusion of income in the list of potential independent variables, the length of time a person spends in a single house may be capturing elements of economic class and age. Given its number of selections, time spent in the same house may also be capturing elements of these two independent variables, among others, better than the variables meant to operationalize them it may be that hypotheses about time spent in a location are underrated, it may also be that this variable is highly correlated with other theoretically relevant predictors. The correlations between the amount of time living in one's house and age is slight, it is significant, as signified by their 0.18 Pearson's R, which was also the case between time spent in a house and income for a value of 0.14. The correlations between these predictors make sense; a stable place to live, while not necessary, assists with the stability needed to attain a substantial income, and time spent in house necessarily correlates with age. The older an individual is, the longer they can spend living in one place, making age necessary, if not sufficient, to stay in one place for some time. The correlation of these two predictors with the amount of time spent in one's house suggests that it is picking up more information than the question's formulation might expect.

3. In addition to determining the average, however, the confidence bands were desired, despite the difficulty of getting a confidence band for a bounded value like the PCP, which ranges between zero and one. For this reason, the confidence bands were determined through bootstrapping; the $2.5^{th}$ and $97.5^{th}$ percentiles of the PCPs captured by the cross-validation were measured, creating a bootstrapped 95% confidence band akin to those created for continuous variables

## Chapter 3 Variations and Alternatives: How BeSiVa the Predictive Perspective Can Contribute to Theoretical understanding

The last chapter demonstrated how the BeSiVa algorithm, and the cross-validatory approach can validate and compare between different theoretical explanations determining what predictors are substantially relevant. By comparing predictive potential, the algorithm compared theoretical explanations of turnout, trying to determine if the choice to vote or not is best

predicted by psychological, sociological, or mobilization theories. It demonstrated that of the three, the sociological approach had the most capable predictors, while also suggesting a considerable role for the strength of party identification, a psychological predictor, as well. This chapter concentrates on more recent incidents, where theory has yet to catch up with the questions, concentrating on guiding theory's development through algorithmic means. This chapter concentrates on two questions: Donald Trump's political rise and the individual perception of the prevalence of minorities, using the algorithm to suggest where theorists should concentrate their efforts.

In the case of the Republican nominee for the 2016 presidential election, little research at the time has been fully explored the likely predictors of Donald Trump's rise to political prominence. To remedy the difficulty of sifting through explanations, Trump's support is explored using an inductive approach to the data at hand, finding places to begin looking for support. Given the recency of the topic, it is difficult to find theoretical work on the origins and support of Donald Trump, and the predictive approach may allow for a better understanding of this unlikely candidate and president.

In addition to tackling a recent problem, understanding the origins of support for Donald Trump, the algorithm and predictive approach also considers an older question, one where the theoretical basis remains precarious. Unlike Donald Trump, the perception of racial minorities, especially in terms of proportions, has been considered in the literature (Alba, Rumbaut, and Marotz 2005). The theoretical background of this question, however, is relatively limited, allowing for the creation of a more comprehensive, theoretically expansive consideration of perception of the proportion of minorities. The results of these two questions are compared and

considered to demonstrate both that theoretical insights may be drawn from predictive methods, allowing for a predictive assessment of theory.

### Some Prior Theory

In their article on the psychological origins of perceived minority prevalence, Alba, Rumbaut, and Marotz focus on two major concepts: the social context of the individual and the affect towards immigration and minorities. The perception of ethnic group size was linked in part to the social context in which an individual was situated, concentrating on the number of minorities that a person lived around. In this case, being surrounded by a greater number of minorities led to the increased belief that minorities are prevalent in the country as a whole. While suggesting that people who lived near minorities would overestimate their prevalence, this part of their theory also implied that minorities would overestimate their own proportions (Alba, Rumbaut, and Marotz 2005). Thus, an individual's perception of the national proportion of minorities was determined in part by the number of minorities in the individual's community, increasing with the number surrounding the person.

Alba et al. propose that an individual's perception of overall minority proportions grows with the number of minorities in the individual's community, but they also concentrate on affect towards minority groups as a predictor. To Alba et al, affect is related to perceptions of prevalence due to the sensitivity of individuals who perceive immigrants or minorities as a threat. The underlying causal mechanism arose from the increased sensitivity that dislike or distrust of immigrants and minorities created. The more that someone felt upset by minorities in general, the more they would notice minorities, leading to an overestimation of overall minority

prevalence. This sensitivity implied that additional dislike of minorities would lead to an increase in the perception of minority group sizes, as an individual whose feelings towards these groups tended towards negativity will overstate their numbers due to a bias based in not wanting any intrusion by immigrants or minorities overall.

Although it is clear how Alba et al. propose context and bias affect perceptions of minority prevalence, their work represents a fertile basis to build upon using inductive approaches for two reasons. First, while Alba et al. have been cited, little work has built theoretically upon their initial offering. What work exists has two contributions, treating Alba et al. as a tangent, or focusing on extending the initial theoretical argument to new contexts. For instance there has been some work drawing on Alba et al. using the concept of innumeracy (Lawrence and Sides 2014), especially focusing on how to correct it, as well as a recent article that discusses the role of social context at the local level (Kunovich 2016). These works, however, tend to either sidestep the theory that Alba, Rumbaut, and Marotz pioneered, or they tend to take it as exactly correct, suggesting a need for further expansion.

In addition to the relatively small expansion that built upon Alba et al.'s original work, there are a few problems with Alba et al. These problems include difficulties in determining exactly what variables the authors used in their data, the 2000 General Social Survey, to estimate perceptions of ethnic minorities. While some of the variables, such as gender, were self-evident, others, such as an individual's residence as urban or rural, were difficult to determine, making the study difficult to replicate and add to the methodological aspect of the argument. In the absence of enough clarity to replicate, and lack of further theoretical development, Alba, Rumbaut, and Marotz's article makes for a promising starting point to develop new theory from an inductive

approach.

In the case of Alba et al, the singular theory and difficulty in replication both suggest a need for additional work, but their meager theory represents more development than the theoretical case for Trump's support. In the case of the 45th president, if there were a selection of theories available, then a deductive approach might be able to determine exactly what led to Trump's support among voters, similar to how turnout in the 2000 presidential election was examined in Chapter 2. While deduction may be thought of as working with a single model and collection of hypotheses, serving as an additional summary statistic on the model as seen later, chapter 2 demonstrated how the use of prediction allows for selection between hypotheses, makes the predictive approach more than another summary statistic. The predictive approach, in conjunction with theoretically driven reasoning, allow for the consideration of different theories in a manner that allows for comparison, but it can also be used to direct theorists' attention in the absence of a strong set of explanations.

Predictive statistics, combined with a deductive approach, best serve our understanding when there are a selection of theoretical explanations, allowing for a comparison of multiple predictors. This was exemplified by the consideration of voter turnout in Chapters 1 and 2. After all, the study of voter turnout has a heritage that dates back to the 1940's, as seen in Campbell et al and the contribution of the Michigan school (1960). From the deductive standpoint, BeSiVa, the predictive approach, and cross-validation all have something to contribute. But in the cases discussed in this chapter, a theoretical basis has yet to be laid down, and political science has either overlooked or only started to understand the underlying causal mechanisms. In order to understand these phenomena, this chapter looks at using the algorithm to speak to new or

underdeveloped problems inductively, beginning with predictive connections and using them to determine what explains the phenomenon. It does so using a slightly altered formulation of BeSiVa, one which can evaluate continuous dependent variables. Although this approach is based in part off predictive sample reuse (Stone 1974, Geisser 1975), Stone and Geisser's algorithm was formulated well before it could be readily applied. Now, the machines have caught up with proposals for variable evaluation and selection to the point where techniques may be easily compared, such an approach may come fully into its own.

## The Deductive Template, and Why New Techniques are Needed

The deductive template, making hypotheses, predicting their implications, gathering data, and testing those hypotheses, may be all about avoiding techniques that eschew theory, but its approach is not as well-followed as the discipline claims. As Yom, citing Jackman (2006), points out, the use of the deductive template appears to be less of a hard and fast rule, and more of a pretension that researchers make to one another. In truth, political scientists often perform inductive studies, and only appear to create a deductive-looking research design at the point of publication, which Yom recommends eschewing, instead advocating for an approach that prizes the deductive component of research, but allows an inductive component to come into play (Yom 2015). What Yom has described is likely to continue, but in the face of difficult studies, the deductive approach's continued utility must be called into question, as it cannot deal with such situations appropriately.

Apart from the fact that the deductive approach to research is not as common as claimed,

as Yom points out, the difficulty with continuing to use the deductive template arises from a variety of different issues related to the current landscape of data. First, there is the question of what to do with the vast collections of data that political scientists are gathering and will continue to gather and have available. These innovations have been especially prominent in the area of textual analysis, where large, overarching theories of such data may be developed (Monroe et al. 2014). But the textual data Monroe et al. describe were developed specifically for the analysis, and it is not difficult to imagine a time when data, developed for uses other than political science, becomes available. When that time comes, the hypothetical deductive approach, already experiencing difficulties (Schrodt 2014) will be fully overwhelmed, and it would be better if an alternative were already available.

The hypothetical deductive, null hypothesis testing approach is useful if the questions asked are not based around the ability to work with more data than an army of theorists could consider. As an example, if a dataset has 1,000 variables, then the ability to make use of the available information is limited by beginning with a few pre-specified hypotheses and only testing them one variable or collection of variables at a time. Combined with the difficulties associated with using null hypothesis significance testing, the hypothetical deductive approach is unable to deal with problems that originate from large datasets. These problems include a massive number of columns, with implications that could keep theorists busy for decades, and a massive numbers of rows overwhelming significance testing, suggesting statistical significance everywhere while providing no substantial findings.

In addition to the question of the deductive approach's utility and necessity for close, pre-specified hypotheses is growing obsolete in the age of large datasets. When a question has only

recently been specified, such as the election of Donald Trump or the theory is sparse, as in the perception of minority prevalence, an inductive approach may supplement prior developments, if they exist. Such an approach allows the data to be pared down, so that the large datasets are trimmed to the point where theorists may begin to work inductively, developing theories around predictors that lead to better models.

While BeSiVa was demonstrated to select variables and create predictive models in the first two chapters, selecting variables and making concrete predictions, it is differentiated from predictive sample reuse by the preservation of a variation on the PCP. In the situation of continuous or linear dependent variables, the predictions are instead based on the closeness of the predicted observation to the measured value in the test set, using a threshold that may be set by the researcher. Such a threshold allows the algorithm's user to determine what constitutes a good prediction, providing greater control than Stone and Geisser's approach.

### Predictive Sample Reuse: BeSiVa's Original Formulation

BeSiVa was developed mostly independently from approaches that came before it. It drew slightly from the writings of Kuhn and Johnson (2013), but was designed around a specific problem rather than from the statistical literature. Despite this situation, the algorithm's roots lie in statistical literature, although it is differentiated by a more intuitive predictive criterion. At a time when computers were incapable of easily calculating the kinds of regressions that BeSiVa uses, well before the development of generalized linear models, Stone (1974) and Geisser (1974), developed predictive sample reuse, an approach that is very similar to BeSiVa. The formulation of predictive sample reuse and BeSiVa differ mainly in terms of the intuitiveness of the results and the regressions at the heart of the data.

There are enough similarities between PSR and the BeSiVa algorithm that it can be considered a precursor. Like the BeSiVa algorithm, PSR is based around the consideration of a test set, keeping data out of the regression to determine the most predictive model given available data. The highly mathematical formulation of this approach demonstrated its potential for variable selection, but its development was hampered by the lack of computational power when it was developed. Despite this, the BeSiVa algorithm has a unique contribution in terms of the intuitiveness of its original predictive criterion, the PCP, something makes BeSiVa's conclusions more interpretable than its predecessor which used the RMSE.

While the goal needs to be to apply algorithmic approaches to problems where they may be relevant, BeSiVa's additional benefit lies in additional intuitiveness inherent in its main criterion. Instead of trying to maximize prediction absolutely, part of the algorithm's charm involves translating machine learning, and predictive analytics into something that can be easily digested and understood. To this end, chapters 1 and 2 demonstrated how the PCP, a bounded random variable that has been previously considered as the overall error rate (Kuhn and Johnson 2013), could be used to select appropriate variables in a logistic regression. Using PCP, and thinking of the variable as a percent, makes it simpler to consider and use in research, which needs to be extended to the continuous case if BeSiVa is to preserve its ability to make more intuitive quantification of predictions.

The PCP, formulated as a percentage, has a definite advantage due to the ease of interpreting it compared to other summary statistics for models. It is especially useful for considering the performance of a model in predicting the dependent variable, due to bounded nature and ease of computation. These similarities ease interpretation due to the scale on which

the PCP lies. PCP is effectively a percentage, running from 0 to 100, and as such is easy to interpret and compare. If a model has a PCP of 50%, then it predicted fifty percent of the test set data correctly. A model with a PCP of 65.7% does a better job of predicting the dependent variable than one with a PCP of 50%, at least for the data that was considered. This benefit would be lost if we returned to PSR's original formulation, due to its reliance on root mean squared error, hereafter the RMSE, to work with a linear regression. For this reason, it would be preferable to alter the predictive criterion, preserving its intuitiveness for continuous dependent variables.

A model's ability to predict may be considered using the RMSE, but there are certain qualities that the PCP has, such as interpretability, that make extending the measure to continuous cases preferable. The RMSE, defined as the square root of the sum of the average residual, as seen in formula 2, and can be used on any continuous dependent variable. When used in a regression table, the RMSE refers to the average difference between the observations used in the regression and their predicted values. The predictive approach, however, calculates this value on a set of data held out for predictive purposes (Kuhn and Johnson 2013), which would be the test set BeSiVa uses. This approach, however, incorporates both the bias and the variance of the linear estimator (James et al. 2013), meaning that it allows for any amount of acceptable noise. Such an approach may work if there is no concept of what makes an appropriate estimator, but the ability to specify the level of acceptable variance allows for a greater conception of the problem, and what makes a good prediction.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n} u_i^2} \qquad (2.)$$

In his discussion of stepwise selection techniques, Harrell (1996) argues that the automated variable selection is typically used to allow researchers to avoid thinking about the problem. Similarly, relying on the RMSE allows for the consideration of how close a model gets on average, but may not be the most appropriate way to determine a model's predictive capability. While the PCP works on categorical data, a reconfiguration that works with continuous data should make the consideration of a problem more intricate, rather than avoiding it, as Harrell suggests. Such a measure forces a researcher to consider the appropriate level of uncertainty in their question, developing a deeper understanding of both the problem and their level of comfort with that uncertainty.

The benefit of extending the PCP to a continuous case involves considering what prediction is close enough for the purpose of a given research question. In the case of the RMSE, there is no right or wrong prediction, only a prediction that gets as close as possible to the measured value. But if the RMSE is used for a continuous dependent variable, the intuitive nature of the PCP is sacrificed; to retain the utility of the PCP, a new means explaining the quality of predictions is necessary. This chapter lays out a similar intuitive measure that explains the predictive capability of a model with a  continuous dependent variable while still falling on a scale from 0 to 100%, making it intuitive than the RMSE

Although the difficulty of understanding the RMSE, by comparison to the PCP, is not that much greater, the formulation of PCP allows for consideration of difficulties that are present within real data. As an example, much of the data that data social scientists work with contains missing values. When the RMSE is calculated, it cannot easily take missing data into account, as it only calculates the deviations between existing dependent variables and the predictions that a

model makes. The RMSE as it is formulated fails to allow for the consideration of how much data could hypothetically be predicted. This would not be a problem if the RMSE were calculated on the data used for the regression for explanatory purposes, but prediction requires an assessment of how well a model works to predict new data. By its formulation, the PCP considers how much data cannot be predicted due to missing values, making it better at assessing a model's predictive power compared to the RMSE.

Although the PCP was designed to work exclusively on categorical data, and was built specifically for binary dichotomous measures, its ability to incorporate information on missing data shows the use of extending it to continuous dependent variables. Part of the PCP's utility is that it can penalize a variable or set of variables if the model that includes them makes good predictions on a tiny fraction of the test data when most of it is missing. By keeping the ratio of correct predictions to total possible predictions, the PCP lets a researcher consider not only whether a variable makes good predictions, but if a variable's missing values negatively impact predictive quality. The RMSE, on the other hand, is incapable of such a comparison, only considering non-missing values and the magnitude of their errors. For this reason, a variation on PCP should be considered for the continuous case, as it provides a way of accounting for missing data as well as the intuitiveness that the PCP provides.

While it is clear that the PCP should be extended to continuous variables, due to its ability to deal with missing data, it is necessary to consider what form such an extension would take. For the PCP and logistic regression, the only question is whether the predictions are the same as the measured values in the test set. Since this is highly unlikely in a continuous dependent variable, any assessment of an extension must begin with difference between the

measured and observed values, the residuals is needed. One way of extending the PCP involving the residuals requires looking at a count of how many predictions are some threshold away from their observed values. This count can then be compared to the number of observations overall, creating a percentage similar to the PCP for a continuous case. Such a measure would be different from the percent correctly predicted as the predictions would only be 'close', rather than correct, for a predetermined definition of closeness. This measure, a Percent Closely Predicted, or PClP, would allow for a preservation of what makes the PCP intuitive but also evaluates a model's predictions of continuous dependent variables.

To allow a researcher to predict a continuous dependent variable, the PClP is proposed as a parallel measure to the PCP in prior chapters. Although it was designed to allow absolute control over what constitutes a 'good' prediction, this strength of the PClP may also turn out to be a weakness. Unlike the PCP, which simply counts the number of times the algorithm correctly classified the dependent variable, the PClP requires a statement of what a good prediction looks like. In attempting to provide more control, this approach requires researchers to operationalize a new concept: what constitutes a 'good' prediction? Allowing for more control may become burdensome, requiring an operationalization of a new way of thinking about residuals, going beyond RMSE where any difference between the observation and the measured value constitutes an error. Although the PClP offers greater control over what constitutes a good prediction, this requires additional consideration, making understanding a good prediction essential.

The PClP's threshold, akin to other tuning parameters, has no mathematically optimal value, making it impossible to claim that some formula may provide the absolute best threshold for calculating the PClP. Some ways to make a better threshold, however, may be deduced from

the PClP's formulation, which allows for a more intuitive use of this new measure. For instance, as the threshold for the PClP decreases, the PClP's value will also drop. This is due to the fact that the threshold indicates a good prediction, and as the range for a good prediction becomes smaller the PClP will be less permissive, causing its value to shrink. Similarly, a large threshold will lead to larger PClP's, as the larger range of values that constitute a good prediction will lead to a more permissive algorithm. While this allows for an understanding of the kinds of predictions a dependent variable might make, it fails to provide a systemic method of calculating and considering differing PClPs. Such an approach is necessary if predictions are to be compared.

$$\frac{|max(DV) - min(DV)|}{c} \tag{3.}$$

As an opening possibility, an intuitive means of considering predictive accuracy arises from scaling the PClP to the range of any given continuous dependent variable. This way of creating a threshold for the PClP allows for the rescaling of a 'good' prediction, on par with the size of any continuous dependent variable. This also enables the same scaling to be used for different continuous dependent variables. Such a formula makes it possible to compare how easy it is to predict different dependent variables reliably, and it allows for the comparison between different dependent variables. An example is seen below in formula 3, where the range of the dependent variable is divided by a factor of *c*. If c were equal to 10, for example, then any prediction within one tenth of the range in either direction will be considered a correct prediction by the algorithm. This is an extremely generous prediction range, but even some dependent variables may not be predicted well, even with such generosity. For these data, a dependent

variable is apparently so hard to predict that a massive range such as the one used with formula 3 (where c == 10) may still not lead to a good prediction.

### The Classification and Regression Tree

While BeSiVa is an approach created for the purpose of predicting values of a dependent variable, and it has been extended to continuous dependent variables with the PClP, it is hardly the only way to predict a continuous dependent variable while finding relevant independent variables. One alternative, the Classification and Regression Tree (CART) allows for the consideration of an excess of independent variables, while also trying to find independent variables that systemically capture how the dependent variable changes. This similarity to BeSiVa made implementing CART intuitive, allowing for the comparison of two separate techniques designed for variable selection and prediction. For this reason, CART is implemented to compete with BeSiVa, providing intuitive, interpretable predictions.

Similar to BeSiVa, CART attempts to predict values of the dependent variable, but it concentrates on splitting the data in a different way, looking for breaks in independent variables that track with divisions in the dependent variable. Taking a dependent variable and an independent variable, CART attempts to divide the variables together in a manner that, by knowing the value of the independent variable, we might also know the value of the dependent variable for the data that are available. It does this by attempting to minimize the sum of squared errors in the split up data (Hastie et al 2008, Kuhn and Johnson 2013). When it has many independent variables to work with, it looks for the variable that divides the data in order to minimize the difference between its prediction and the observed value of the dependent

variable,in this case,  the sum of squared errors (SSE). After finding this variable, it continues to divide the data with it, before looking for other independent variables that might continue to divide the dependent variable based on the independent variables' values. Through this approach, CART attempts to divide the data up in a manner that leads to the best possible prediction of the dependent variable.

Dividing the dependent variable along splits in available independent variables, CART predicts of the dependent variable while also selecting the most useful independent variables to make that prediction. This means that like BeSiVa, CART conducts variable selection, excluding irrelevant variables from its results (Kuhn and Johnson 2013). Once it has figured out which splits are most relevant for predicting the dependent variable, CART creates a set of decision rules making it possible to use a set of independent variables to determine the final value of a dependent variable. These rules are just a set of if then statements that lead to a set of final values, called terminal nodes,which may be thought of as the final predictions made by CART.

While CART allows for the consideration of independent variables to model a dependent variable, its concentration on fitting just one dependent variable makes it vulnerable to overfitting. The results of CART are known as regression trees, and use a complexity parameter to prune the trees, eliminating splits that are unlikely to translate from the model into prediction. This is done by increasing the size of the sum of squared errors, based on the number of different predictions the model can make. The number of different predictions, referred to as terminal nodes, represent the complexity of the regression tree. Like a regression with too many independent variables, a regression tree with too many terminal nodes will not make good predictions on a new set of data. This complexity parameter has no mathematically optimal

value, but it is possible to determine a good value for the parameter from cross-validation, allowing for a more predictive regression tree.

The benefits of CART lie in the results that it creates, which are more intuitive than advanced approaches such as boosted regression trees or bootstrapped aggregation (Kuhn and Johnson 2013). The larger number of variables it selects and its more involved approach to variable selection, however, mean that its results are decisively harder to interpret than BeSiVa's. Despite this, CART's relative ease of interpretation, and intuitive means of dealing with missing data make it an attractive alternative to BeSiVa for making predictions. For this reason, CART is both introduced and used as an alternative to BeSiVa, making comparisons between the two techniques on two questions: Donald Trump's support and the perception of minority group sizes.

### An Empirical Example

But before getting too involved with the inductive approach, let us begin with a brief demonstration of the difficulty of using a small selection of variables for making predictions, focusing on the Trump feeling thermometers. For this demonstration, we shall use the feeling thermometer for Trump, as well as a series of preselected independent variables. As an initial determination of support, 5 independent variables are used: an individual's age, minority status, gender, party identification, and ideology, using numeric operationalizations of the last two variables. These variables may seem like reasonable predictors of support, but when the RMSE is gathered for held out observations, it is clear that a different approach is necessary to predict Trump's support.

Table 1: An example fo a model to predict support for Donald Trump. This model, using a few variables to predict affect shows repeated examples of statistical significance, and a large R squared.The RMSE and PClP, however, show that the model could be improved in its ability to make predictions.

| | |
|---|---|
| Constant | $-20.344^{***}$ |
| | (3.621) |
| Age | $0.238^{***}$ |
| | (0.054) |
| Minority | $-2.104$ |
| | (2.153) |
| Male | $3.206^{*}$ |
| | (1.806) |
| Party ID | $6.408^{***}$ |
| | (0.551) |
| Ideology | $7.604^{***}$ |
| | (0.996) |
| Observations | 1,067 |
| $R^2$ | 0.361 |
| Adjusted $R^2$ | 0.358 |
| RMSE | 29.329 |
| Held out RMSE | 29.58 |
| PClP | 21.42% |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

When examined from an explanatory perspective, the example appears to do a good job of modeling Trump's support, but more work is needed to improve the model's predictions, as seen from its summary statistics such as the RMSE. The RMSE is gathered using the validation process on the model that appeared in chapter 2. This process involves holding out 20% of all observations, selected at random, and using the held out observations to determine how well the model makes predictions. This is then compared to a regression that has all included data to determine how well the model predicts Trump's support. While the model itself, seen in table 1, shows significance for all variables except minority status, the model's RMSE tells a different

story, a story which is verified by the validation process.The RMSE on the full model is 29.33, suggesting a poor fit. For the observations in this model, the average error is larger than a quarter of the range of the dependent variable, which runs from 0 to 100. This means that if an observation were selected at random, it would be difficult to determine whether someone liked or disliked Donald Trump and how strongly they felt about him from this model. Similarly, the root mean squared error of the held out data is 29.58, a value that suggests that on average, predictions will fall so far from the dependent variable that the model could not reasonably be claimed to make an accurate prediction.Specifically, the errors are so large that given a new observation, it would not be possible to predict whether someone supported Donald Trump or not, or whether their support or lack thereof was strong. In attempting to predict support for Trump, variables that seem reasonable for describing an individual's level of support fail dramatically in this example.

But as described previously, the RMSE may not be the most appropriate way to capturing the quality of predictions. If instead, the data were compared using the PClP, as described above, it might suggest that despite the large average, there are still a sizable proportion of good predictions. To do this, the process of holding out observations and predicting those observations is repeated, gathering the PClP. The PClP is as described in formula 3, with 10% in either serving as the threshold for a good prediction. When the PClPs are gathered with the same divisions of data and independent variables, however, the results are not encouraging. The mean PClP of the 100 runs is a paltry 21.4%, with a median of 21.7%. Such an approach suggests two things. First, that while the PClP was designed to minimize difficulties with the RMSE, there is an agreement between the two approaches on this particular question. Second, the agreement is

that these variables are not capable of predicting support for Trump in a systemic fashion.

While the demographic variables are unable to predict Trump's support with any reliability, it may be that such demographics are not useful predictors of overall support. But it is also possible that their low predictive power is a symptom of Donald Trump's unprecedented candidacy. In this case, one of the models related to demographics might serve to better model the support of Hillary Clinton or Jeb Bush. This may be tested, however, using the same model with a different dependent variable. To test this possibility, the same cross-validation approach used to generate the PClP in table 1 was rerun, using the feeling thermometers for Jeb Bush and Hillary Clinton. Just like with Trump, however, Clinton and Bush's feeling thermometers are not predicted well by the variables, suggesting that the difficulties in predicting Trump's support lie somewhere besides his unconventional appeal.

Table 2: The results of the empirical test of the feeling thermometer, using age, minority status, gender, party id, and ideological identification. These summaries of the PClPs show that the feeling thermometers are difficult to predict using an empirical test with conventional variables of candidate support, and conventional candidates. With such difficulty predicting candidate support using conventional variables of support, it is difficult to justify using the feeling thermometers as something that may be predicted and used for drawing inferences, despite the attempts to do so (Enders and Smallpage 2016).

| | Jeb Bush | Hillary Clinton | Donald Trump |
|---|---|---|---|
| Minimum | 16.25 | 20.00 | 16.25 |
| First Quartile | 19.17 | 23.75 | 20.00 |
| Median | 20.83 | 24.79 | 21.67 |
| Third Quartile | 22.50 | 27.08 | 22.92 |
| Maximum | 27.50 | 31.67 | 27.50 |
| Mean | 21.00 | 25.30 | 21.42 |
| Standard Deviation | 2.28 | 2.32 | 2.24 |
| Variance | 5.21 | 5.36 | 5.01 |
| Skewness | 0.24 | 0.33 | 0.15 |
| Kurtosis | -0.08 | -0.30 | -0.22 |
| Missing Values | 0.00 | 0.00 | 0.00 |
| N | 100.00 | 100.00 | 100.00 |

Despite rerunning the same test with different candidates, the feeling thermometers for

more conventional figures barely improve the results of the predictive tests. The results of predicting the feeling thermometers for Bush and Clinton, shown in table 2 alongside the results of Trump's PClP values, suggest that even if the dependent variable featured a more conventional candidate, the variables do not lead to a better prediction. In fact, using the same independent variables to model Jeb Bush's feeling thermometer leads to a worse prediction on average based on table 2's values of the PClP. With the same independent variables, Clinton's feeling thermometer has an average PClP of 25.3%, with a 24.8% median. Jeb Bush, meanwhile, has a PClP of 21% on average, with a median PClP of 20.8%. Despite their supposed appeal to different groups in a more systemic fashion, neither Bush nor Clinton's feeling thermometer is predicted much better by these variables. The conventionality of the candidate does not appear to matter, and may even hamper predictiveness of feeling thermometers, suggesting that the problem may be more systemic than Trump's unprecedented candidacy.

There are several conclusions that may be drawn from this example. The first involves the possibility that Trump and the other candidates' support are not based primarily in these questions, and that it would be better to search elsewhere for predictors of candidate support. If a well developed theory of Trump's support existed, that theory could be tested using this approach, placing relevant variables into a model and doing explanatory and predictive tests. Predictive testing would allow for the consideration of the utility of those theories, but the recency of the Trump phenomenon means that such a theory does not yet exist, providing an opportunity for the predictive approach. In this circumstance, a predictive approach may be able to guide theorizing by enabling the in-depth exploration of the data at hand, finding relationships that might predict support for a candidate.

## The Datasets

In order to understand the support for candidates such as Donald Trump, especially in the early portion of the 2016 campaign for president, it is necessary to access data that focuses on Trump as a candidate specifically. To this end, the American National Election study provided a dataset that includes such questions on Trump as a candidate, concentrating on Trump as a primary contender in its 2016 Pilot study. Some of these questions compared Trump to other candidates, asking respondents whether they preferred him to other contenders in the republican field, as well as whether they preferred him in a head-to-head competition with likely democratic contenders. In addition to these questions, the ANES asked respondents how much they liked Trump directly using a feeling thermometer. Capturing large quantities of demographic and political data, the ANES provided a survey that allowed for the assessment of Donald Trump in comparison to other relevant variables.

While the ANES' pilot study was necessary for understanding the sources of support behind Donald Trump, following the perception of the number of minorities required a different data source. The data used by Alba et al. for their original study (2005), the 2000 edition of the General Social Survey was employed. The GSS in 2000 asked a series of questions concerning an individual's beliefs on the proportion of minorities in the United States, which Alba et al. used to advance their own theoretical arguments. Similarly concentrating on the individual, the GSS asked about the prevalence of minorities within the United States, as well as the number of minorities in an individual's community, among other variables. In order to understand an individual's perceptions of minority prevalence, the same dataset that Alba, Rumbaut, and

Marotz used to capture it originally was reenlisted for the purpose of testing BeSiVa, in concurrence with CART.

### The Dependent Variables

Determining sources of support for populist candidates requires considering different ways that individuals might support a candidate. After all, an individual may support a candidate for president, but would prefer that another candidate did better in the primaries. The individual may seek to vote for a primary candidate even if they seem like an unlikely pick for the eventual nomination. Trump's position was the latter when the ANES survey was taken, but understanding support for a particular candidate is something that needs to be considered in order to model or predict such support.

Trump's candidacy was considered in several different questions throughout the ANES. Questions concerned whether Trump was the respondent's preferred Republican candidate, as well as who respondents would support in a head to head contest with Donald Trump and the Democratic candidates at the time. In order to understand support for Trump, however, a question was needed that spoke both to understanding who truly liked the eventual president as well as the potential of the algorithm. The feeling thermometer was chosen as a means of avoiding the possible considerations of who was electable, concentrating on how people felt about Donald Trump to understand his eventual election. Given the difficulty in predicting it using the more conventional variables, it also suggested that a different selection of variables one that the algorithm might provide, would be preferable in predicting Trump's support.

$$Minority\ Perception \equiv \ln\left(\frac{African-Americans_{US}+Hispanics_{US}}{Whites_{US}}\right) \quad (4.)$$

In addition to predicting support for Donald Trump, the algorithm was also applied the the perception of minority presence in the United States. While it was possible to use the raw percentages a respondent estimated for these groups, Alba et al's dependent variable was recreated, which can be seen in formula 4. The recreated dependent variable allowed for direct comparisons between the results the authors obtained and the variables selected by the two algorithms. To Alba et al., minority perception is a function of the percentage of African-Americans and Hispanics the respondent estimates to be in the United States, divided by their estimates of Whites. While this variable was not on the same scale of the feeling thermometer, it was possible to compare its results to those in Alba et al. directly, and an appropriately chosen PClP allowed for comparison to the feeling thermometers as well.

### The Independent Variables

In an inductively based study comparing the results of two algorithms, the independent variables must be the same, to enable the comparison of results, and to make sure that the results could be interpreted. As an example, the 2016 ANES pilot study asked a set of questions that had the potential to explain Trump's support among the voting electorate, including questions about party identification, ideology, and anti-establishment sentiment. However, a number of columns of data were eliminated, due to the difficulty in relating their content to the dependent variable. These columns concentrated on the mechanics of the pilot study, specifically the amount of time that respondents spent on each question in seconds, as well as whether the respondent needed prompting to finish a particular question. Originating from the online survey's timers and reminders, no clear path to a theoretically satisfying answer appeared to exist from these questions, making it sensible to remove them from consideration.

In addition to questions relating to the mechanics of the online pilot study, a series of questions that were conceptually identical (with minor differences) to the dependent variable were removed from consideration. These questions included the respondent's preference among all republican candidates, as well whether respondents preferred Republicans over Hillary Clinton, akin to the choice that voters would later make on election day. Although they were originally included, the questions were removed in later analyses, due to their similarities to questions concerning Trump's overall support. At the time the survey was taken, January (ANES 2016), these questions served as a less descriptive feeling thermometer, seeing whether specific republican candidates were preferable to Hillary Clinton. Given Trump's support among Republicans, as opposed to independents and Democrats, the idea that such a variable would be useful for predicting support seems unlikely, as it appeared to merely be a restatement of the dependent variable.

In both the ANES and GSS cases, the variables were pared down to a selection that catered to the strengths of the methodologies. Due to the likelihood of slowing CART to a crawl, any variable with more than 20 unique cases was removed from consideration, while variables that were of low variance (Kuhn and Johnson 2013), defined as possessing more than 95% identical values, were also kept away from consideration. To make BeSiVa's operation more likely to be fair, any variable that consisted of over 50% missing data was also eliminated from consideration. With these elements used to standardize the independent variables that could be selected from each dataset, the algorithms could compete on fair ground, each having the same information for both dependent the independent variables.

## Methods

Similar to the last two chapters, the datasets was divided into several subset for differing purposes, but given that the measure for determining predictive quality, the PClP, differed from PCP in a substantive way, the remaining data were divided into third sets. Two of the divisions, the training set and the test set, did not change in their use, but a third subset of data, the validation set, was kept away from BeSiVa's optimization and CART's regressions, based on the recommendation of Shmuelli and Koppius (2011). While the previous tests of BeSiVa always concentrated on the performance of the test set, the algorithm has changed due to the new predictive criterion for evaluating a continuous dependent variable. In this case, the reformulated BeSiVa is no longer concentrating on a measure that has been discussed in prior work, such as Kuhn and Johnson (2013), warranting a more stringent test of its predictions. The PClP, though similar to PCP, represents an innovation in considering what constitutes an error, and therefore calls for a more stringent consideration of what constitutes a good prediction.

Given that BeSiVa has been reformulated to use PClP, rather than PCP, checking on the quality of the predictions becomes more essential than in the last two chapters. A question of whether BeSiVa was merely optimizing on this data set stayed open, and using a validation set strengthens the idea that BeSiVa makes relevant, substantively significant predictions on new data. While prior analyses did not include a separate validation set, fewer divisions are desirable due to a problem of data resolution. When there are few observations, as seen in previous chapters, then divvying the data into many different sets for testing purposes becomes problematic. Small test sets, however, make ties more likely, making it harder to determine the superior predictor. When PCPs or PClPs are identical, the BeSiVa algorithm stops, meaning that

a smaller test set, shrunken further by additional divvying out of data, makes the algorithm more likely to end in a tie. It is preferable to avoid divvying out additional subsets, diluting the resolution of the data and making prediction harder, but it is also necessary to verify the quality of the PClP, and to guarantee that the BeSiVa algorithm and CART are competing fairly.

Despite the difficulty that data resolution poses to the question of making the best prediction, and allowing the algorithm to run freely, it is necessary to add a separate validation set. Such a dataset will strengthen the criteria that judges the PClP, and to make a fair competition between BeSiVa and CART. After all, BeSiVa has had the chance to pour over the test set data; it has optimized its prediction on the test set, and in this comparison, BeSiVa has a distinct advantage in prediction that CART lacks if their performance on the test set is compared. A separate validation set requires the two algorithms to predict a dataset that neither has seen, allowing them to compete in predicting data without giving either an unfair advantage. Thus, the setup of a separate dataset from prior data subsets, the validation set, is necessary. Despite the loss of data resolution, making ties more likely, BeSiVa and CART may be tested with a dataset that makes the competition between the two algorithms fair. This fairness is accomplished by checking not only how the algorithms predict on data that neither have seen or used, putting the algorithms on equal footing in a test of their predictive capabilities.

Having parsed out the data into these 3 sets, training, test, and validation, both the BeSiVa algorithm and CART were allowed to make predictions on the validation and test data, using their respective approaches to prediction. They were trained on the same set, and each was allowed to predict the test and the validation sets using the same criterion, the PClP, with a threshold calculated from formula 3. To ensure that the results of an individual division of data

were not a fluke, this process of modeling and predicting was repeated 100 times. Once the algorithms had run and made predictions, results that allowed for a comparison between the two algorithms were collected.

Once the algorithms had run, several pieces of information relevant to comparing CART and BeSiva's performance were collected, starting with measures of variable importance. While CART provides more information on importance than BeSiVa, any variables considered important by either algorithm were included and saved for consideration. This may be seen as limiting to CART, as it could hypothetically select a variable many times, but consider that variable of low importance. By comparison, BeSiVa selects its preferred variables in terms of their inclusion or exclusion in a final model. Despite this potential limitation of the design, CART's selections should still ultimately tend towards variables that it considers important, even if they were only saved based on a binary criterion. This is due to the way that variable importance is calculated, which is based on how much the variable lowers the criterion used to generate CART's regression trees (Kuhn and Johnson 2013). Therefore, if a variable tends to make the model better (by making the splits in the data resemble the dependent variable, CART's criterion), it should show up repeatedly in different runs of CART, making it more important. For this reason, the recommendations that CART and BeSiVa made concerning variables of importance were saved, allowing for an equitable comparison between the recommendations the two algorithms made.

In addition to capturing the variables of importance, the PClPs of each run were also saved. Both CART and BeSiVa generated two sets of PClP, one for the test set, and one for the validation set. While a PClP is not necessary for CART's operation, it proves useful for

comparing the predictions that each makes. These results were saved as percentages, and form an essential part of the results, as they allow for a direct comparison between the algorithms. Having saved the results necessary to make that comparison on both the GSS and ANES datasets, the algorithm's results could be considered, allowing for better understanding of what makes predictions on perceptions of minority prevalence, and Trump's support.

It may appear that a comparison between the two questions, minority perception and Trump's support, is impossible due to their differing scales and necessarily different PClPs. To continue guaranteeing that the variables were on a level playing field formula 3 was used to generate a threshold of the PClP that was comparable between the two questions. This meant that a fifth of the range for Alba et al's. dependent variable was used as a threshold, making sure that the variables were compared in an equivalent manner. Choosing a threshold similar to the one used for the feeling thermometer enabled a comparison of these two dependent variables, allowing a comparable consideration of the ability to predict them in an inductively oriented study.

## Results

Having detailed how the results were gathered, CART and BeSiVa were both run on the same datasets, and their results were considered for the questions of interest, starting with the perception of minority prevalence. Figures 1 and 2 compare the selected variables for predicting perception of minorities using Alba, Rumbaut, and Marotz's constructed dependent variable, as seen in formula 4. They show any variable selected more than five times for BeSiVa and 25

times for CART, respectively, due to CART's far larger selection of variables. The first thing to note concerns the number of variables selected. True to its attempts to deal with missing data, CART selected a much larger number of variables. This is due in part to its attempts to compensate with missing data, adding variables used in constructing surrogate splits to deal with missing observations. These variables required a paring, but the results of CART suggest that there is some concurrence with Alba et al.'s theory.

Figure 1: Variables selected by BeSiVa 5 or more times for predicting perceptions of the proportion of minorities. As Alba, Rumbaut, and Marotz suggested, perception of number of minorities in the community and social matters (2005), but social context may be extended quite dramatically.

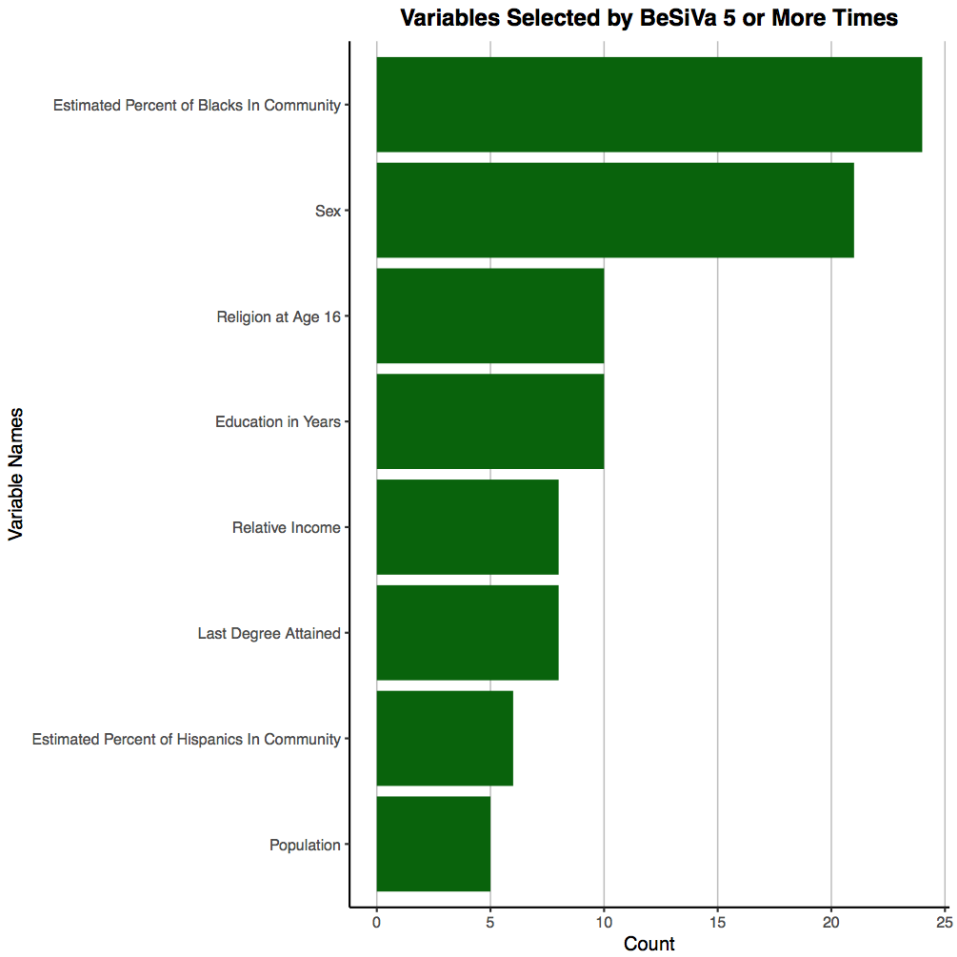
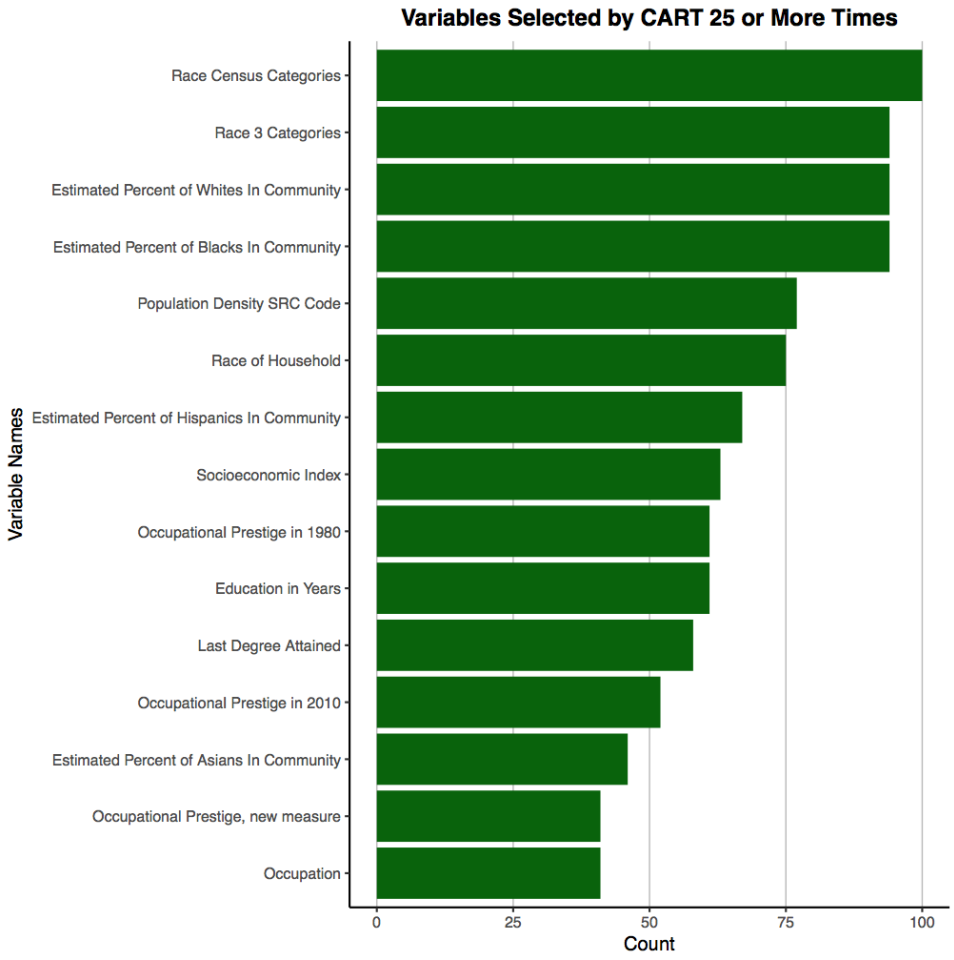


Figure 2: Variables selected by CART 25 or more times for predicting perceptions of the proportion of minorities.CART's expansive selection makes it difficult to interpret the results to allow for induction, but it is clear that community and contextual factors are necessary to predict the perception of minority presence from CART's results.




Validating Alba, Rumbaut, and Marotz (2005) CART's selections tended towards racial variables, focusing especially on an individual's race or racial status, as well as the racial status of the community that surrounds them. Estimated percentages of blacks and whites in the community, as well as hispanics, are among CART's most selected variables, as are identifications of race based on census data and the racial makeup of the household in which the individual lives. Similarly, questions of whether someone lives in a rural or urban area based on

population density prove relevant for CART's predictions, as do sex and education. When CART selects variables, it does so very broadly, but the variables that it prefers appear to correspond to theory.

For the most part, BeSiVa concentrates on variables similar to CART, but its selections are more judicial. BeSiVa selects far fewer variables than CART, both in terms of the number from which it chooses and the number of times it selects them. BeSiVa's selections, though fewer, are also more consistent, partially due to its use of listwise deletion, partially due to its tendency to cease operations if a tie is discovered. Both of these aspects of BeSiVa's operation will lead to a smaller selection of variables if there are only a few observations. Despite the differences in BeSiVa and CART's selections, the number of selections each makes, and the variables that are chosen, the two approaches also share similar a focuses in the variables that they choose.

While BeSiVa is far more conservative, both in terms of the number of variables selected and the diversity of the selections, the variables it prefers have much in common with those selected by CART. For example, the number of African Americans in the community is the most selected variable, and the number of hispanics arises frequently as well. Race is also selected regularly, although BeSiVa does not find it as relevant as CART, and racial variables appearing in a smaller proportion of the algorithm's runs. BeSiVa, however, prefers to concentrate on demographic variables in the GSS, and its second most preferred variable was the individual's sex, followed immediately by education. These racial and demographic variables suggest a general accord with the theory, but there are some notable variable selections that suggest places where the theory needs extending.

Despite generally agreeing with Alba, Rumbaut, and Marotz (2005), both BeSiVa and CART have notable differences on what predicts perceptions of the United States' ethnic makeup. One of the major deviations, which both consider in some capacity, is the religious makeup of an individual's community. The religious aspect is selected using two variables, concentrating on religion at the time of the survey and religion at age 16. Religion, however, is selected as an essential variable in over a third of cases for CART, and 10% of the time according to BeSiVa. CART was more focused on the religion of the individual at the time the survey was taken, while BeSiVa focused more on what a religion was practiced in an individual's house at the age of 16. While Alba, Rumbaut, and Marotz discussed the context in which an individual lived, religious belief observances were not considered in their discussions of social context, an oversight which both BeSiVa and CART were able to correct.

In determining what predicts an individual's perception of minority proportions, it was clear that religious belief mattered in the case of Alba et al.'s dependent variable. The algorithms demonstrated that the original study overlooked aspects of social context that might better explain how an individual perceived ethnic minorities' presence. The social context was considered an aspect of an individual's perception of minorities, and the lack of a religious variable in the original study was one example of an overlooked part of that context. Similarly, however, economic variables such as satisfaction with one's financial status proved essential in understanding what drives the perception of the number and percentage of ethnic minorities. These results suggest a fulfillment of what Alba, Rumbaut, and Marotz indicated, even as they differed from the original theory.

Despite the discovery that an individual's concerns about their financial status, their sex,

and their religious belief all matter, Alba et al.'s theory only needs slight extensions to account for these new predictors. For instance, religious belief ties very closely into the social contexts that Alba et al. used to explain perceptions of minority presence. Some religious practices are more racially diverse than others (Dougherty 2003), varying the exposure to racial diversity in an individual's life and therefore their perceptions of minority presence. This unconsidered aspect of social context extends to an individual's finances and socioeconomic class, making these variables more of an extension of the the theory rather than a divergence.

In addition to the question of why religious belief might affect an estimations of minority prevalence, both algorithms also selected socioeconomic status and financial comfort, elements of an individual's life that Alba et al. did not consider. Socioeconomic status and financial comfort, however, tie closely into economic class, another element of a person's social context. It is well known that race and class are often tied closely together, and the discomfort with financial status and socioeconomic variables would also tie into social context. Individuals of lower class may be exposed to fewer minorities, but they can not choose to change their location, and thus the number of minorities, and social context, is fixed. For this reason, these results are not truly that shocking, as both an individual's religious beliefs and their socioeconomic status may tie back into their social context, a part of Alba et al's original argument.

While it is possible to imagine religion and class as components of an individual's social context, it is difficult to explain the fact why is clearly a predictor of the perception of minority prevalence. While sex's selection may seem counterintuitive, part of individual's perception of minority prevalence is based in the amount of threat they feel from ethnic minorities, according to Alba et al. Despite research suggesting that that sex matter when considering immigration

policy (McLaren 2003) men and women may perceive immigration differently, especially in terms of threat. Although an individual's sex may not appear to be a predictor of the perception of minorities and immigration, it may also tie into a perceptions of immigrants as a threat and therefore the perception of ethnic group size.

Figure 3: Variables that CART selected for predicting the Trump Feeling Thermometer in the ANES 2016 pilot study. CART Selects a much larger list of variables, making it difficult to draw inductive conclusions, but President Obama's performance, as well as concerns over who can handle refugees and immigrants (syrians_a, best1, ISSUES_OC14_7) are of high priority to CART's considerations.

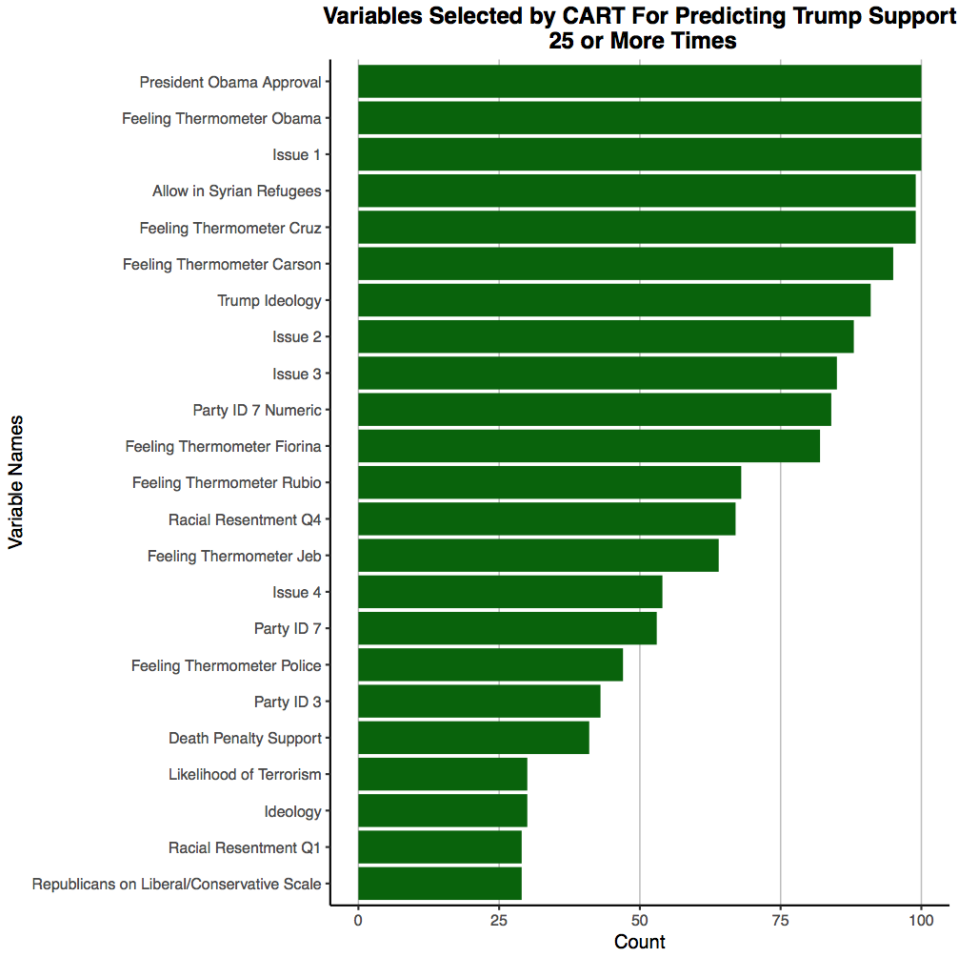
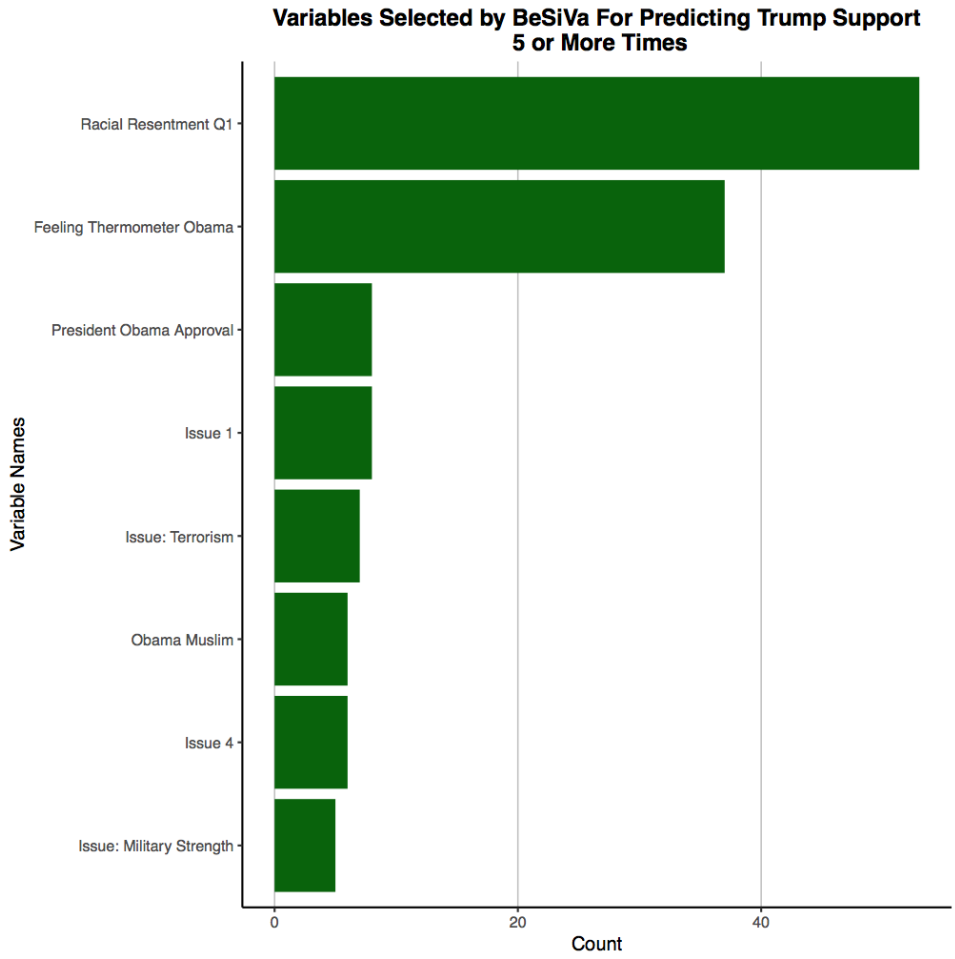


Figure 4: Variables BeSiVa selected for predicting the Trump Feeling Thermometer in the ANES 2016 pilot study. Racial resentment is the best predictor of support for Trump, followed closely by a respondents feelings about Obama and concern about terrorism, suggesting a referendum on racial issues and concerns about the administration preceding Trump.




Having explored the results from the two algorithms related to Alba et al's findings, let us now turn to the results of the tests on the ANES data, looking at what predicted support for Donald Trump. As seen in figure 3, CART's results point towards the idea that the approval of Trump was a referendum on Barack Obama's time in office, repeatedly selecting variables that

concentrated on feelings towards Obama. These variables included assessments of how the president was doing his job, as well as a feeling thermometer of Obama directly.This was also important to BeSiVa's assessments, as seen in figure 4. BeSiVa selected a different variable as its first choice, but the two concurred that feelings towards Obama mattered heavily in predicting individual's preferences towards President Trump.

While feelings towards Obama mattered quite a bit in assessing Trump, the algorithms also agreed that issues mattered in determining how people felt about Trump. The matter of which party was better at handling a specified issue was a concern to both BeSiVa and CART, each of which placed the issue most important to the respondent, referred to as issue 1, near or at the top of variables for predicting Trump's support. While BeSiVa had some agreement as far as competence and representativeness was concerned, it was also more interested in the issues that were selected, looking especially at terrorism and security, Military strength, and women's rights and immigration. The immigration and terror variables were selected by CART, although it was more concerned with the evaluations of Obama, but their concurrence suggested that some issues, such as terrorism and security, matter heavily in evaluations of Trump.

BeSiVa somewhat concurred, but of its 100 runs, the algorithm was far more interested in one of the racial resentment variables that were available to it. It's most commonly selected independent variable concerned the racial resentment of respondents. Representing an implicit condemnation of African-Americans, the question compared the Black experience to other minority groups, which were well known to have previously condemned and integrated into society. Asking whether the respondent agreed with the idea that these groups overcame prejudice without looking for special favors, the question elucidated on the racial bias the

respondent held. BeSiVa selected this variable in over half of the runs, suggesting its concern with racial discrimination as a predictor affect towards Trump. In this case, the inductive result is obvious, Trump's rhetorical appeals to racial animus attracted individuals to his side and cause, increasing their affect towards Trump.

While BeSiVa seemed to believe that racial animus was most essential in predicting Trump's support, CART considered voters to be more concerned with the issues that Trump promised to focus on. In both BeSiVa and Cart's opinions, however, a feeling thermometer of Obama followed the first variable directly, suggesting that when it came to feelings towards Donald Trump, assessments of the president were a necessary component in predicting the feeling thermometer.

While CART and BeSiva had some concurrence with what variables are necessary to predict Trump's support, both algorithms had a great deal of difficulty predicting who was a supporter. This can be seen in the feeling thermometer predictions. As seen in figure 3.6, all predictions related to feelings towards Donald Trump are remarkably poor. The PClPs however, were exceptionally low, falling within the 25% range for both algorithms. The final results of the ANES feeling thermometer for Donald Trump suggest that the collection of variables available to both BeSiVa and CART remained unable to explain individuals' affect towards Trump. This fundamental unpredictability suggests one of several possible outcomes, few of which are encouraging for the use of feeling thermometers as a dependent variable.

### Why are feeling thermometers so hard to predict?

While it was not difficult, using the criterion laid down in formula 3, to predict the perception of minorities, the feeling thermometers proved very difficult to accurately predict. Figure 5 shows the results of the 100 runs to predict Alba, Rumbaut, and Marotz's dependent variable, and while the predictions are hardly as accurate as those on the choice to vote in chapter 2, it is still possible to at least predict a majority of observations accurately. By comparison, predicting support for Trump, as seen in figure 6, is exceptionally difficult. In the first of several explanations of the poor outcomes for the feeling thermometers, it is possible that the poor predictions arise from the possibility that the algorithms, CART and BeSiVa, fail at feature selection, despite arguments to the contrary (Breiman et al. 1984, Rogers 2014). Despite this possibility, there are several reasons why the failure is unlikely to lie with the algorithms. First, the two separate methods would have to both fail simultaneously, and a viable alternative would need to be demonstrably preferable. This was not the case.

Figure 5: The results of the percents closely predicted on the validation set, which neither CART nor BeSiVa saw. While it is clear that CART makes a better prediction than BeSiVa, BeSiVa makes a comparable prediction with far fewer variables.

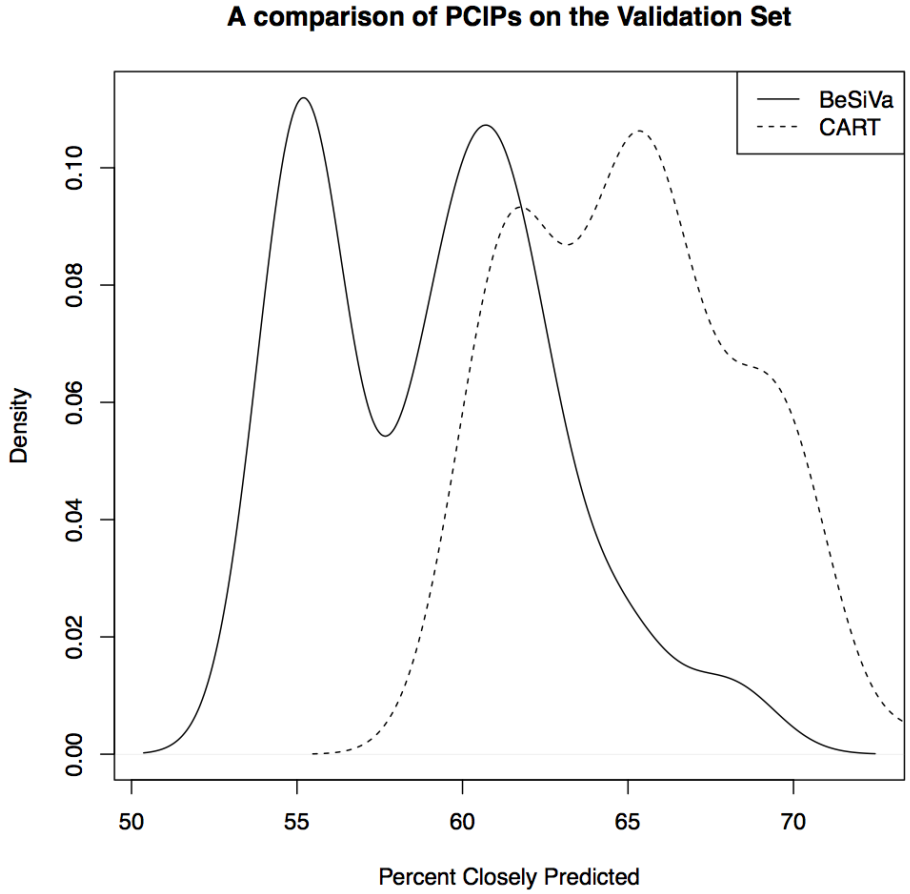
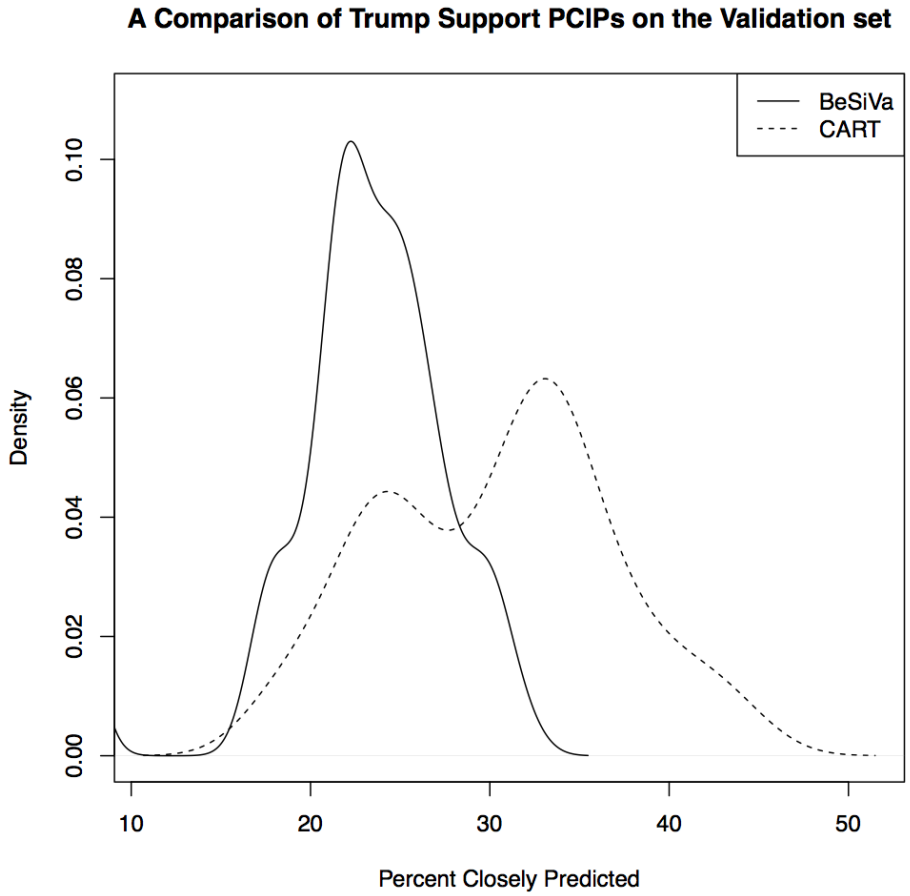


Figure 6: A comparison of the PClPs on the validation data for the ANES feeling thermometer for Donald Trump. While CART once again outperforms BeSiVa, BeSiVa's list of results are more intuitive. Sadly, neither approach was very useful for predicting Trump support, suggesting the difficulty of predicting feeling thermometer results.




In order to dismiss the use of two predictive algorithms, it would be preferable to show that another approach was better at predicting the dependent variable. There should be a comparable empirical example, such as the one described in the condemnation of the hypothetical deductive approaches. To demonstrate that the problem lies in the choice of algorithmic approaches such as BeSiVa and CART, the empirical example should be capable of doing a better job of Modeling affect towards Trump, and making predictions that explain the

affect towards Trump. It would be preferable that there was a demonstrable way of accurately predicting the dependent variable, even if it was not algorithmic. Sadly, the empirical examples also only made bad predictions, as seen in table 2, suggesting that the problem does lies in the dependent variable rather than the methods.

The empirical example of Trump support suggests that the problem is with the dependent variable. This is verified by the results from Alba, Rumbaut, and Marotz's data, which backs up the idea that these algorithms may make accurate predictions . With PClPs in the 60% to 70% range on the validation set, as seen in figure 3.5, the case of perception of minority prevalence shows that the algorithms are both capable of performing excellently with a large threshold for closeness. The threshold for PClP has the same standardized range for feeling thermometers and perceptions, equivalent to 20% of the range (or 10% of the range in each direction). While a more rigorous threshold might create a more parsimonious predictive model (albeit one with lowered PClPs), the same threshold makes comparisons between different dependent variables possible. The results of those comparisons show that despite the difficulty in making predictions on the feeling thermometer, CART and BeSiVa are still capable of predicting a dependent variable well. Although the dependent variable differs, the use of the same threshold allows for the consideration of whether there was a simultaneous failure of both approaches in predicting affect towards Trump. It appears that instead, there is some difficulty related to the available data, rather than the approaches that were used to try and predict the dependent variable.

Having discussed why CART and the BeSiVa algorithm are unlikely to have simultaneously failed, it is necessary to consider alternative explanations for the difficulty in predicting affect towards Trump. A second possibility is that no combination of available

independent variables could model the feeling thermometer for Trump with a degree of accuracy. It may be that the Trump feeling thermometer represented affect towards a true political outlier, someone whose political career could not be predicted by conventional means. If that were the case, however, then affect towards more conventional candidates, such as Hillary Clinton or Jeb Bush, would have been captured with less marginal accuracy. In an attempt to determine whether this was the case, the dependent variable of Trump was substituted for Hillary and Jeb in an approach akin to the empirical example, and the results are displayed as a part of table 2. Given that they similarly fall well below any reasonable threshold, the summary statistics for these PClPs demonstrate that the problem is not just with Trump, but has something to do with the feeling thermometer.

Having discussed why the feeling thermometer is likely to be the problem, rather than the algorithms, or the candidate, it is necessary to consider what might make the feeling thermometers problematic. One possibility is that the relationship between the relevant independent variables and dependent variable is not linear. This is nearly comforting, suggesting that a more advanced approach, one concentrating on capturing non-linear relationships between variables, might be able to model feeling thermometers. This would require placing a more flexible method at BeSiVa's center as a way of circumventing the nonlinearity. One option would be to remove ordinary least squares and substitute in generalized additive modeling, which uses splines to model different functional relations between the dependent and independent variables (Wood 2006). Such a model, however, would add an additional layer of complexity to BeSiVa, and leave CART unable to predict the dependent well. Since the goal of BeSiVa is to allow for interpretability of the results and make good predictions on the dependent variable, it does not

seem that such an approach would be desirable

Having considered a number of possibilities for why the feeling thermometer may not be easily predicted, the results suggest that the problem lies in the way that the dependent variable is measured. On the one hand, rating individuals on a 0-100º scale might suggest capturing additional information, but the large number of choices may mean that individuals' true preferences are lost within a sea of noise. In this case, the dependent variable cannot be modeled due to poor measurement, as none of the provided variables can predict the feeling thermometers variation with any degree of accuracy. While a more intricate technique such as generalized additive modeling might be able a better prediction, it appears likely that the feeling thermometer simply represents a poor means of capturing an individual's feelings overall.

### Inductive Explanations?

After considering the variables that BeSiVa selected, it appears that the algorithm selected variables that aligned closely with the theory that Alba, Rumbaut, and Marotz suggested. After all, the number of African Americans and Hispanics within the community are both considered highly relevant, as are gender and education. Notably, however, BeSiVa also seemed interested in religious background, as well as an individual's satisfaction with their finances. These variables seem to align well with the social context aspect of the original work's theory. There is, however, a risk that the theory is not laid down with enough care, and that the resulting variables may have suggested a theoretical capriciousness that Shapiro (2002) warned against. In this case, each new variable or concept is easily incorporated into the preexisting structure of Alba, Rumbaut, and Marotz's theory. Its generality, while not intended, may have

proven to be too wide to be of use.

While Alba, Rumbaut, and Marotz's consideration of social context as a concept that predicts perceptions, the risk is their theory is too broad, allowing for the incorporation of any demographic variable. If this is the case, then the theory they created is far too capricious to be useful, as per Shapiro, who condemned theoretical capriciousness. After all, the use of social context might imply that any demographic information is relevant to perception of the proportion of minorities. Alternatively, however, it is possible that the authors are unaware of the possible implications of their own theory, especially elements surrounding class and religious belief. In this case, an individual's socioeconomic status could be a key driver of predicting minority perception, possibly through a correlation with communities of color and class. In this case, the class of an individual predicts their contact with minorities on a daily basis, making individuals of lower SES more likely to overestimate minority prevalence. For this reason, the selections of income related variables may be able to extend the theory set out by Alba, Rumbaut, and Marotz, serving as key elements of the social context that were overlooked in the original theory.

The possible interplay of class and racial communities suggests a place for variables such as financial satisfaction and socioeconomic status. Despite this contribution, it does not explain the primacy of religious belief in BeSiVa's variable selection, or CART's willingness to entertain the idea of religion as important to predicting perceptions of minorities. In this case, religious belief might be a sign that individuals raised in a nontraditional religious background are more likely to perceive racial prominence differently. Religions such as Hinduism or Buddhism may also correlate to a different racial background, or a different racial background in the communities that the individual inhabits, making a religious community essential to

understanding the social context influencing a person's perception of minorities. These beliefs then, like an individuals' financial satisfaction, suggest an extension of the original theory which was found algorithmically, suggesting a role for CART and BeSiVa in extending theory.

While it was possible to extend Alba et al.'s original theory via predictive approaches, it is difficult to imagine that either the predictive or empirically based approaches are useful for determining what predicts support in feeling thermometers. Regardless, the predictive approaches' suggestions should be considered as an opening point for determining affect, especially once a more reliable dependent variable for measuring affect is constructed. CART and BeSiVa differ in their consideration of the most important variables for predicting Trump's support, but their suggestions show some concurrence, especially in terms of racial resentment, issues of import, and evaluations of the president before Trump.

The first, and most obvious place for constructing inductive results on Trump arises from the BeSiVa algorithm's favorite variable, but aspects of the same concept crops up in CART's selections as well. This concept is operationalized by BeSiVa's selection of racial resentment, which is the most commonly selected variable among those that were available. If BeSiVa is to be believed, people supported Donald Trump due to the racial appeals that he made during the campaign. There was some suggestion of racial animus as one of a series of factors that led to Trump's rise (Schaffner 2016, Carmines, Ensley and Wagner 2016) and it appears that racial resentment plays an important part in predicting whether an individual felt positively towards Trump.

Despite Schaffner's suggestion that sexism was also a factor (2016) no variable that indicated animus towards women was found by BeSiVa in over five of its runs, and as such do

not appear in figure 4. While variables related to discrimination against women are a part of CART's results, such as whether there is discrimination against women, and opinions on equal pay, which can be seen in figure 3.5. These variables, however are selected less than a quarter of the time by the algorithm, compared to variables related to racial animus and affect towards Barack Obama. While some measure of sexism may have been present in the 2016 presidential campaign, appeals based on gender do not appear to dominate in predicting affect towards Donald Trump, especially compared to racial appeals.

While Schaffner explores racial animus as a measure of support for Donald Trump, it is clear from his informal work that a confluence of factors contributed to affect towards Trump, and BeSiVa and CART concur on that point. While BeSiVa suggests that racial resentment is key to understanding what made people like Trump, it also appears that their feelings on several issues predict his support. These issues, highlighted by BeSiVa and CART's selection on who would best handle these issues, show that voters cared about who would allay their concerns over matters of policy. But which policies are most likely to predict support? When policies come up in the predictive models, the most selected ones concern a mixture of immigration and terror, suggesting a concern with security that is seen in both algorithms' results.

Although Carmines Ensley, and Wagner suggest that Trump's support arises from populism, the mixture of social conservatism and fiscal liberalism that they use to define the term is not relevant in predicting whether Trump is liked, while looking at issues that touch on racial concerns, such as terror and immigration, help to predict Trump's support with some accuracy. It appears that through these issues, voters who support Trump register levels of anxiety with the current state of security, both in terms of borders and terror, looking for someone they believe

can provide that security. Trump's support then, also came from issues that related especially closely to border and overall security, suggesting that his concentration on simple policy prescriptions were relevant in obtaining people's affect.

In addition to racial animus and issues related to safety, the algorithms suggested that evaluations of Trump were based on evaluations of the administration that preceded him. In this case, the focus on feelings about Barack Obama, and evaluating how he did as a president were relevant predictors according to CART and BeSiVa. Like the example with Alba, Rumbaut, and Marotz's theory, this too ties into prior work on how individuals feel about a candidate. Usually, however, retrospective voting is considered from the perspective of national elections, however, rather than primary contests (Fiorina 1981). Predicting affect towards Trump, however, requires considering how individuals feel about his predecessor, including a feeling thermometer of Obama and evaluations of how Obama was doing as president. There may have been a racial component to this, given that BeSiVa and CART both selected a question concerning whether Obama was a muslim in some of their runs, but the evaluations of the past President were always highly preferred variables in the algorithms' selections. It is therefore necessary, from the predictive algorithmic perspective, to assess support for Trump based on support for Obama.

While the appeal of Trump may have initially appeared confusing, it is clear from both algorithms that a series of variables predict Trump's support better than others. It makes sense then, from both BeSiVa and CART, to conclude that there are strong explanations for why people liked Donald Trump. The strongest predictors, according to both algorithms, include retrospective evaluation of President Obama, racial appeals, and concerns related to specific issues, especially immigration and terror. It may be that a better dependent variable would have

resulted in effective predictions, but the algorithmic results are still capable of predicting Trump's support among primary voters, and suggesting predictors that correspond to prior work, prioritizing some over others.

### But why BeSiVa?

Through the BeSiVa algorithm, it was possible to develop a series of inductive explanations for the candidacy of Donald Trump. These explanations tended to center around a variety of independent variables, but the main takeaway from the method involved the difficulty in making predictions for feeling thermometers. Despite a generous threshold for what constituted a 'good' prediction, people's preferences for or against Donald Trump were hard to predict, and nearly identical to the PClPs in the empirical example. This difficulty suggested that even with these predictive algorithms, affect towards the future President could not be gleaned from the methods with the available list of variables. The results, however, also suggested that even if the predictions from all approaches were poor, CART's predictions were the least poor. CART's ability to make better predictions continued to be true in the alternative example, the paper by Alba Rumbaut and Marotz

Unlike the Trump Feeling Thermometer, both BeSiVa and CART were capable of making good predictions on the question that Alba, Rumbaut, and Marotz considered, to the point where an inductive explanation was completely possible. BeSiVa was useful in considering the perception of minority presence, despite its results' similarities to the theoretical proposals of Alba et al. In this case, however, the theoretical explanations on which Alba et al. concentrated, social context and exposure to minority communities, were reinforced. Most notably, the

algorithms suggested that religious belief and background may represent aspects of an individual's social context that the authors simply neglected. CART largely agreed with these predictions, but its lists of variables were far more elaborate, concentrating on the individual's racial background more than the background of their communities. In this case, however, these considerations led to a better prediction overall, and CART's more complex means of determining which variables mattered made the better prediction of the two approaches for the problems that Alba et al. considered.

When it came to the validation set, data which neither BeSiVa nor CART had seen, but needed to predict, it was not possible for BeSiVa to create a better prediction than the one that CART made. Even though the two approaches were comparable in terms of their predictions on the test set, CART outperformed BeSiVa very slightly on that data. When this was extended to the validation set, it was clear that CART was better at making predictions on completely new data. The competition on the validation set, as seen in figures 5 and 6, was won by CART, where the predictions that CART made were more accurate based on a set of data that neither approach had previously been allowed to use.

Despite its ability to make consistent predictions, as seen in Chapter 1, BeSiVa's first contest against an alternative approach demonstrated that such an approach was capable of making better predictions on new data. BeSiVa, however, outperformed CART in another more important area: the intuitiveness of the provided results. Where CART provided an better predictions, it did so at the cost of a less intuitive description of what makes one variable more important than another, and included a massive list of results. By comparison, BeSiVa gave an intuitive evaluation of the independent variables via thresholds that may be grasped without

CART's more elaborate approach to making predictions. While the results BeSiVa provided were less predictive overall, BeSiva compensated for this lack of predictive capability by the comparative intuitiveness of its results and recommendations.

BeSiVa's results, while slightly less predictive than a more intricate technique like CART, are also far more intuitive, and CART leans more towards intuitiveness than other predictive approaches. Techniques' complexities range from simpler approaches like CART to 'black boxes', such techniques require only data to generate good predictions, but they fail to make it clear how those predictions were generated. While CART looked for whatever variables decreased the error in its split, BeSiVa attempted to determine whether a particular independent variable should or should not be included in a regression model. BeSiVa based this decision on how well a model made with that variable predicts a set of additional data, held out for predictive purposes. While CART focuses on how much a variable improves its sum of squared errors (Kuhn and Johnson 2013), BeSiVa was trying to improve a percent of observations that fell close to their measured values, according to the threshold laid down in formula 3. BeSiVa lies further on the intuitive side of making predictions than CART, and can serve as a bridge between more complex techniques and the approaches that political scientists know and use regularly.

While it would be nice to be able to gain both intuitiveness and maximize predictive quality, the nature of prediction requires a tradeoff between these two desirable qualities in any technique. A technique may be able to predict extraordinarily well, but if it fails to make those predictions intuitive, then understanding the relationship between its selected predictors and the dependent variable could hardly be incorporated into a scientific approach. With BeSiVa, an explicitly prediction driven technique, the results may not be the most intuitive, but a decrease in

intuitiveness allows the technique to become more comprehensible, making it easier to use and incorporate into a scientific approach.

## Conclusions

BeSiVa is an optimizing method, looking for patterns within the data that make the best prediction on a given dependent variable, and it can be used in different ways to develop new knowledge. This chapter featured BeSiVa with an inductive approach to determining answers to two questions: an individual's perception of the proportion of minorities and support for Donald Trump as a political candidate. Although these two phenomena might seem disparate, they are connected by the approach that was used to weigh their theoretical bases, the BeSiVa and CART algorithms. The results that BeSiVa and CART demonstrated suggested that it was possible to extend theory, in the case of Alba, Rumbaut, and Marotz's article, and to generate an understanding of the limits of certain kinds of measurements, in the case of the Trump feeling thermometer.

In order to consider these two questions, BeSiVa was extended to model continuous dependent variables by creating a predictive criterion akin to percent correctly predicted for OLS. This measure was created by using a new way of considering what constitutes a good prediction, the percent closely predicted, or PClP for short. Defined as the percent of observations that fall within a certain threshold, using formula 3 as a way of generating standardized thresholds, the PClP allowed researchers to define what constituted a good prediction. Although a comparable threshold was used for both cases, it was also clear that only one of the two could be predicted with any degree of accuracy.

With a permissive threshold, it was possible to develop models with highly predictive elements, as demonstrated with Alba, Rumbaut, and Marotz's article on individual perceptions of the proportion of minorities (2005). It was clear that the theory Alba et al put is supported by a predictive method, but that the results of CART and BeSiVa demonstrated where the theory could be extended. While it was expected that the results of testing Alba, Rumbaut, and Marotz's dependent variable would extend the theory in new directions, it was clear that the theoretical approach was capable of predicting perceptions of minority proportions well. The algorithms demonstrated Alba et al.'s theory by extending the social context which a person inhabits to include religious belief and background, as well as social class. The inductive explanation from these results suggest that certain religious backgrounds include greater racial integration, which affects the social context and the background against which an individual perceive minority prevalence.

The algorithms' findings suggest that the theories, which Alba et al. proposed were capable of being extended, and that their theory was generally correct. It is possible that Alba et al's theory, as Shapiro (2002) warned, was too capricious, failing to truly explain what predicted the perception of minorities. But Shapiro was concerned with directionality, whether the concepts that theories used could be pinned down and compared, in terms of increases and decreases in the relations between concepts. It is clear from their article that Alba, Rumbaut, and Marotz's relationship, that social context and increased contact with minorities increase the proportion of perceived minorities in the country. For this reason, the theory was verified and extended inductively, suggesting new directions to explore the role of social context in individual perceptions.

One of the reasons that Alba et al.'s theory may be inductively extended lies in the fact that through the PClP, it was clear that excellent results could be obtained. By comparison, some dependent variables are not so easily predicted with available data. This was demonstrably the case with Trump and the other candidates' feeling thermometers in the ANES data. Despite selections of variables such as racial resentment, evaluations of Barack Obama, and issues relating to terror and immigration, it remained extremely difficult to predict support for Trump with any accuracy. Indeed, the means of the PClPs were far lower for the feeling thermometers, and no approach was able to recreate the predictions seen in previous chapters. Given BeSiVa's success up to this point, and the similar failure to predict other candidates, as seen in the empirical example seen in table 3.2, it appears that feeling thermometers themselves are difficult to predict. While it is possible that a less intuitive approach might be able to predict affect for candidates, the question of whether it is worth sacrificing the ability to understand how the prediction is made remains open.

In addition to considering predictive approaches to two new questions, this chapter also explored a new division of data than as a means of testing predictive quality more stringently. This was necessary due to the new reconsideration of the PCP with a different threshold, the PClP, leading to training set, which is used for regression in BeSiVa and CART, a test set, which is optimized by BeSiVa, and a separate validation set, which provided a more stringent test for BeSiVa, and a test for CART, as CART does not divide the data in the same manner. A validation set is also helpful for comparing predictive capabilities, as CART doesn't provide a way of determining the quality of its predictions (Kuhn and Johnson 2013), like the PCP or PClP in BeSiVa. This separate set, along with the similar thresholds allowed the two approaches to be

compared. Even though the BeSiVa algorithm was unable to beat CART for predictive quality, it is clear that the use of a predictive approach, whether aiming for intuitiveness or maximized prediction, can provide additional theoretical insight, extending theory in directions that were not previously considered.

# Chapter 4: Concluding Remarks

## The Origins of BeSiVa

This dissertation arose from a challenge that was pointed out in my graduate course on research methods. This challenge, the idea that we could not use or consider every available predictor, was most likely to be meant as as a statement of fact, a fair assessment given the reasoning that lay behind it. With that in mind, however I saw what our professor had stated as a challenge to overcome, rather than a factual statement. The fact that all data could not be used in a research design was justified theoretically, as theory provided a direction to the methods that would be used. But what if the theorists were missing something important? While techniques such as ordinary least squares could still provide unbiased estimates while missing variables that were important to answering the research question (Fox 2008), the results would remain incomplete. Being in the middle of things, however, the idea of creating something that would allow me to consider the whole of the data, communicating with theorists on an equal footing, was postponed.

While I recognized and continue to recognize the necessity of theory, the idea of using all available data in holistic approach to answering research questions remained something that I wished to explore. This idea became more relevant for the problems that I found myself dealing

with while working at a private firm, coming up with the algorithm in response to multiple questions related to data that I encountered working for political campaigns. These questions included what to do with the excess of data that parties were able to provide, as well as the fact that my employer had some seemingly idiosyncratic demands relating to how models were assessed.

The algorithm was first conceived of based on two unusual requests that my employer at the company wanted me to incorporate into the research approach that I was using. These included taking of some fraction of the data out of the regression and modeling, and finding a way to use that data to assess the model's predictive capabilities. For the employer, it was 10%, and similar fractions were used throughout the dissertation due to their connection to cross-validation approaches (Kuhn and Johnson 2013). In addition to the extra assessment, in what I'm sure was supposed to be a helpful move, some clients were also able to secure large collections of available data related to contacted voters. As such, it seemed like a waste to simply work with the same predictors as dictated by theory, especially when that theory came up short. The demands for a held out subset of data, and the opportunity presented by the additional predictors led to the creation of BeSiva, which was designed to meet these requests.

Despite BeSiVa's origins outside the political science literature, the literature demonstrates that explanatory statistics remain dissatisfying as a means of settling research questions. What Schrodt provided in his fiery condemnation of international relations (2014) was merely the tip of a long and mostly ignored tradition of searching for a better way to verify findings. Works such as Cureton (1950), Meehl (1954), and others in the psychological literature saw the need for techniques that completely sidestepped concerns related to explanatory

statistics. While techniques like predictive sample reuse could sidestep some of these concerns (Stone 1974, Geisser 1975), they were ignored, either due to concerns of data mining (Bartels 1997), their similarities to problematic approaches like stepwise regression (Harrell 1996), computational limitations, or simple ignorance. Meanwhile, scholars such as Cohen (1994) and Gill (1995) sought to avoid explanatory techniques such as significance testing, due to their clear misinterpretation in practice, and the difficulties in conveying them to others (Haller and Krauss 2002). The BeSiVa algorithm, and the predictive approach more generally, allows for a clearer, more intuitive means of testing research questions without the problems that the explanatory approach has accrued.

### Differences in Opinion

While it allows for a more intuitive means of considering research questions, the primary contribution of the BeSiVa algorithm is an opinion based in prediction rather than explanation. This opinion does not concern what explains variation in the dependent variable, the main objective of the explanatory approach (Shmueli 2010), but what explains variation in the dependent variable for a new set of observations. As Shmueli elaborates, this leads to a different set of priorities, which may seem confusing at first. The priorities, however, are more in line with the goals of the average social scientist in some ways, especially given the difficulties seen in using explanatory statistics properly.

The predictive approach's concerns with the quality of predictions over simplicity and explanatory power may seem counterintuitive, but the difficulties in following best practices with the explanatory approach suggests a need for alternative, more rigorous techniques. For

instance, the consideration of model assumptions when the variable is dichotomous, rather than just running logistic regression, remains absurdly rare, despite Achen's (2002) warning that logistic regression is overused. According to Achen, the overuse of the logistic regression estimator did not take the dependent variable's real underlying probability structure into account, suggesting that logistic regression is dramatically overused. This may be considered more closely with BeSiVa, and the predictive approach. Specifically, if logistic regression works well when used to predict data that were not included in the model originally, then a more sophisticated estimation technique is not necessary.

In addition to concerns related to the use of logistic regression, BeSiVa allows for a closer consideration of the need for parsimony. Ignoring Achen's thoughts on logistic regression, researchers also rarely care about Achen's rule of three, which states that any model with more than three independent variables was poorly specified. While this may be true, chapter 2 provides a possible refutation to this rule of three. A model may be misspecified, as Achen puts it, but may still do a better job of making predictions with more variables. Prediction also provides a guide for parsimony; a model with too many independent variables may make a worse prediction, allowing for a statement of how many independent variables are too many, contradicting Achen's opinion. The predictive technique allows for a better sense of the number and selection of variables that best model the dependent variable, more so than the explanatory statistical approach that is commonly used.

Despite the fact that the approach has been heavily disputed as a means of testing hypotheses, political scientists continue to use explanatory statistical approaches, even with the severe difficulties in implementation. Beyond the problems of data mining, deliberate and

accidental (Gelman and Loken 2013), further problems persist. Another major difficulty concerns the fact that it is unlikely that anyone has ever modified their p-values to appropriately reflect the number of models that were created in addition to the ones that were published, as recommended by Reinhart (2015). The pure explanatory approach's utility has been disputed since the 1950's in psychology (Meehl 1954), but political science continues to use it. Schrodt's response to these failures was to leave the discipline behind entirely, leaving the whole of academic social scientific research behind (2014). Researchers who stayed, like Gelman, seem to be angling for a new kind of approach; Gelman favors Bayesian techniques, as does Silver (2012), but this approach remains controversial (Gelman 2008). This dissertation set out to advance an alternative approach to answering research questions in the social sciences, one which demonstrated that a predictive technique was capable of providing new insights. Through prediction, it is possible for political science to cease relying entirely on explanatory insights, and to begin a recreation of political scientific findings from a predictive perspective. Prediction sets us up for a new way of considering hypotheses, especially old ones, allowing for a reconsideration of major findings, with an eye towards more utilitarian considerations. The predictive power of a given model or variables from either a cross-validatory or algorithmic perspective provides the ability to quantify our uncertainty about the utility of that model, and to be ready for the next set of challenges in understanding politics.

What this dissertation provides then is nothing less than a way to remake and reconsider vast swaths of the political science literature from a perspective that eschews explanatory perfection in exchange for substantive insights. With the techniques that were discussed and explored within each chapter, especially the BeSiVa algorithm, the predictive perspective is

considered, using the algorithm as a way of exploring how prediction might enable further consideration of different hypotheses. This perspective allows for a reconsideration of any finding that was tested and validated using the explanatory approach, with the goal of validating or invalidating it through a predictive criterion.

### BeSiVa, Briefly

But in order to discuss the major findings of the chapters, it is important to consider the main tool that enabled their discovery. The best subset in validation algorithm (BeSiVa, for short) was developed to solve the problem of an excess of independent variables with a given dependent variable and to make sure a test set, a subset of the data that was not used in the regression, was well predicted. Only requiring a dependent variable, a list of independent variables, and the data to be used, BeSiVa begins by taking the dataset and dividing it into two separate subsets used for different purposes, as seen in figure 1 in chapter 1. These sets are referred to as the training and test set, based on the terminology established by Kuhn and Johnson (2013); the training set will be used to make models, while the test set will be used to determine model quality. Having prepared the data to find the best subset of predictors, the algorithm begins validating models to find which independent variables lead to the best possible prediction.

Once the data are divided into these two subsets, the algorithm begins searching for the independent variables that make the best prediction on the dependent variable, at least for the given test set. The algorithm begins by taking all available independent variables and making 1 variable models, one for each independent variable. It then tests these models' predictions by predicting the dependent variable, using the data in the test set to avoid overfitting the model

(Clark 2004, Kuhn and Johnson 2013). The quality of each prediction is determined by a different predictive criterion, the PCP for binary dichotomous dependent variables, and the PClP for continuous variables, the ratio of correct (or close, in the case of the PClP) predictions to the total number of possible predictions in the test set. Once it has determined which model maximizes the PCP, the algorithm saves the independent variable from that model, and repeats the process, keeping that variable in all future models, which is repeated until a tie occurs, or the (user specified) maximum number of variable is reached. Through this process, BeSiVa allows for the consideration of a larger selection of independent variables, maximizes prediction on a separate set of data held out for that purpose, and predicts a given dependent variable.

### The Replication Addendum

Having described the algorithm, as well as reiterating the need for predictive techniques, let us delve into some of the ways that the predictive approach might be adapted for more general use. Seen in the empirical example in chapter 3, taking a single model and cross-validating it should allow for an understanding of not only whether the results of a study are true from the explanatory perspective, but also how useful they appear to be (Shmueli 2010). If the cross validated PCP (if the dependent variable is categorical), is lower than one over the number of categories, then the model is not particularly useful for making predictions, and another means of exploring a given phenomenon should be considered (Kuhn and Johnson 2013). Despite its apparent ability to correctly explain variation in the dependent variable, the model is incapable of making any assessment of observations that it has not previously seen.

In attempting to cross-validate a single model, it may be possible that the model is

excellent at predicting, or it may lead to extremely poor predictions, but such a determination remains to be seen. Using cross-validation on older models is referred to as the replication addendum here due to the possibility of reconsidering many findings the political science literature, with an eye towards prediction rather than explanation. Instead of simply replicating prior work, the use of a cross-validation statistic enables a new way of determining the utility, as well as the explanatory capability of a model; the model's ability to predict the situation, as well as explain it, is considered separately. The likelihood of future replication may be assessed (Hindman 2015), as well as the utility of any given model, through the use of cross-validation techniques. Even without BeSiVa, it is possible that the predictive approach may benefit future research through cross-validation, which by itself is capable of increasing the rigor with which a model is tested, providing a separate assessment of theoretical utility.

### Deductively Chosen Variables

While it is possible that the predictive approach may be considered and used to test the veracity of a single model, using the BeSiVa algorithm specifically provides new insights about the comparable relevance of different theories. Through BeSiVa, prediction can weigh a large selection of independent variables and assess their relevance as predictors, in the literal sense of the word. As described in chapter 2, the predictive approach can be combined with a selection of deductively chosen variables, based on different theoretical answers of the same research question. As an example, chapter 2 weighed three different approaches to the question of the choice to vote. These approaches, grounded in different theoretical traditions, were compared by taking the variables which each technique recommended, and providing them to the algorithm.

The algorithm then selected which of the given variables were most capable of predicting the dependent variable, before assessing how the model made with those variables would perform.

Having successfully ranked the utility of variables that hailed from different theoretical traditions, the algorithm's results were interpreted and used in a model. To create this model, the variables were added to build the most predictive model possible from the algorithm's recommendations. Through variable selection and predictive techniques, the theories were allowed to compete, and the most useful theoretical conceptions of vote choice were added to a single model. Included theories were compared and contrasted, and in building the most predictive model, their utility was weighed. Some theories were found wanting, such as the mobilization model and the relevance of finances, while others appear to be absolutely unconquerable, but the results suggested the substantial relevance of each variable that was considered.

The BeSiVa algorithm, in conjunction with a set of predefined theories, allows for the comparison and consideration of the relevance of each theory while providing a direct counterargument to points that Achen made concerning model specification. Recall that to Achen, any more than 3 independent variables led to a poorly specified model, and there was a tendency to overuse logistic regression (2002). As demonstrated by the algorithm, a model may be poorly specified, but this does not preclude its utility. A model with somewhere between four and six predictors provided a better prediction than one that had only three independent variables, and one that had many independent variables, as seen in figures 5 and 6 in chapter 2. In addition to the ability to compare several theories, BeSiVa demonstrated that even if a model was not specified according to Achen's theoretical critiques, it was still capable of being

extraordinarily useful.

Through the deductive use of BeSiVa, it was possible to refute prior methodological critiques in the literature, such as Achen's, and it was possible to determine which theories of turnout were more useful for determining turnout decisions, at least in the 2000 presidential election. For instance, elements of the demographically oriented sociological theories of turnout, as well as the psychological theories of voter turnout heavily outperformed mobilization theory. The mobilization theory pioneered by Rosenstone and Hansen ([1993] 2003) was effectively eschewed by the algorithm, treating the variables meant to assess network placement as nearly irrelevant. Meanwhile, a supposed essential component of socioeconomic status, financial indicators, dropped out of the algorithm's results. Without any input other than providing relevant variables, this discovery backs up one of the findings of Teixeira, that money only makes a difference as a threshold; once an individual is financially comfortable, financial status becomes irrelevant to political participation (1987). Through BeSiVa, it was possible to determine that when predicting vote choice, some theories are more useful than others in a given context.

Thanks to BeSiVa, it was possible to determine the primacy of elements of sociological and psychological theories of the choice to vote in the 2000 presidential contest. But the power of the algorithm is that it may attempt to test theoretical predictions not only cross sectionally, but across different time frames as well. As an example, perhaps the predictors that BeSiVa selected in chapter 2 are only relevant in the context of the 2000 American presidential election. The implication for future work is clear; would predictors such as financial status and party contact matter more in the 1994 election? In 1980? The power of a predictive approach is the

ability to determine the relevance of a given theory in incredible detail, allowing us to seek out when it becomes useful or ceases to help in understanding a concept of interest. While the replication addendum shows that the predictive approach can be incorporated seamlessly into a deductive research design, BeSiVa's ability to select variables allows for the consideration of multiple theories at the same time, and determining multiple theories' overall relevance to answering a research question through their predictive capabilities.

## The Inductive Approach

While it is possible to use the BeSiVa algorithm deductively, choosing between a few potential predictors that have previously been theoretically verified, the algorithm's success at discovering theoretically relevant predictors has also been verified. Seen in Chapters 1 and 3, the inductive approach was first used to test the algorithm, determining whether the variables that it selected made sense, preferably agreeing with available theories. In the case of the choice to vote, the algorithm's predictions were surprisingly well aligned with prior work in the literature, even if the theoretical support in that work was somewhat lax. BeSiVa agreed entirely with the idea that the choice to vote could be seen as habitual; when prior vote was provided, BeSiVa selected only that variable in 97 out of a hundred different runs, despite the other 655 independent variables available to the algorithm. In all cases, the algorithm selected prior vote, a variable that had a rich history within the literature (Brody and Sniderman 1977), showing the power of habit as a way of determining who would or would not vote. An attempt to use the algorithm to inductively determine what predicted whether someone voted or not ran directly back to a predictor that was well known within the literature, suggesting that induction with the

algorithm would be possible for newer questions.

Having demonstrated that BeSiVa came up with good selections for a theoretically specified question, even when irrelevant variables were provided, it seemed reasonable to try it on new questions with less precedent in the literature. In the cases in chapter 3, the algorithm provided results that demonstrated an expansion, rather than a refutation, of an underdeveloped theory. The first attempt to use the algorithm inductively arose from Alba, Rumbaut, and Marotz's work, which suggested that an individual's social context was essential for determining their assessments of minority prevalence tin the United States. By exploring a large body of variables, it became clear that Alba et al. had overlooked relevant aspects of an individual's social context, and that what they had overlooked mattered quite a bit. Alba et al believed social context mattered based on an individual's situation in their community, whether they lived in a rural area or were of foreign birth, which affected minority exposure and therefore their assessment of minority proportions (2005). The algorithm, however, demonstrated that Alba et al. had overlooked essential aspects of the social context, such as religion and elements of one's financial status. These predictors suggested that there were elements of an individual's exposure to racial minorities that were not included in the original assessment of the theory. They also demonstrated that while essentially correct, the extension of Alba et al's theories was not only possible but necessary.

In addition to considering individual perceptions of minority prevalence, the algorithm was also applied to feeling thermometers, specifically the feeling thermometer for Donald Trump in the 2016 pilot for the American National Election Study. Through analyzing Trump's feeling thermometer in the ANES, the algorithm determined that racial resentment was a key predictor

of individual's feelings towards this unlikely candidate. It also demonstrated that Fiorina's (1981) theory of retrospective voting was essential when considering support for Trump, given the relevance of assessments of Obama's presidency. Through the BeSiVa algorithm, it was possible to develop theoretical starting points for understanding the Trump phenomenon, despite its difficulties in predicting feeling thermometers more generally.

Although it provided insight on predictors that allowed for considerations of Trump's support from a theoretical standpoint, BeSiVa also demonstrated the difficulty in predicting candidate affect using feeling thermometers. In this case, the feeling thermometer was difficult to predict exactly, with large variations between observed and predicted values leading to low PClPs. While it might be that BeSiVa's predictive capability was overrated, the use of a second algorithm known as CART suggested that the problem did not lie with the predictive technique. Rather, it appeared that the algorithms, BeSiVa and CART, were doing a poor job of predicting support due to questionable specification of the dependent variable. The large errors suggested that in this case, the feeling thermometer provides an illusion of additional specificity. Through the combined effort of two predictive algorithms, it was clear that the feeling thermometers were leading to the loss of individual's true feelings about Trump and other political candidates. Through the use of two predictive algorithms, it was possible to demonstrate that there were ways to inductively consider new research questions, enabling an algorithmic development of starting points for theory on new questions.

The inductive approach demonstrated that through BeSiVa, it was possible to make inductive determinations about ways that different theories could be extended or developed. This was shown by looking at Alba, Rumbaut, and Marotz's question of individual assessments of

minority prevalence (2005) as well as the Trump feeling thermometer in the 2016 American National Election Study. With Alba, Rumbaut, and Marotz, it was clear that their theory was capable of providing an overarching understanding of what led to assessments of the prevalence of minorities in the United States. There were, however, some notable oversights, especially in terms of religious belief, which may change the racial makeup of an individual's social context, as well as economic class, which changes the ability of an individual to select their social context. With Trump, it was clear that individuals were looking at the last President of the United States very closely when deciding their feelings on the eventual president, as the algorithm repeatedly chose assessments of Barack Obama as predictors of Trump's feeling thermometer. But there were also suggestions of a racial aspect to Trump's appeal, as the main predictor that BeSiVa chose focused on racial resentment, with sporadic selections of racially tinged issues such as immigration and terrorism. Through all of these assessments, however, it was possible to see how BeSiVa selected predictors that led to different theoretical paths worth pursuing exclusively using the predictive approach.

### Future Directions for BeSiVa (Car in the Garage)

Having discussed ways that the algorithm's findings were used to develop around theory, it is also worth considering a few changes the algorithm needs to ensure its continued utility, as well as improvements. With this in mind, some of the future work demands that BeSiVa be made more researcher-friendly, preferably making the algorithm faster and more computationally efficient. At this point, the algorithm's operations are relatively slow, likely due to its reliance on R's basic regression commands. If the algorithm were implemented using matrix algebra functions to calculate regression coefficients, eliminating the additional difficulties that remain a

part of using the algorithm, its results might be gleaned far more quickly.

In addition to trying to make the algorithm faster, one of the major difficulties in working with BeSiVa has been figuring out how to improve upon its current operation. One of the main options to make the algorithm more thorough increases the rigor of BeSiVa's validation scheme. With its single test set, the risk of identifying an unrepresentative relationship between variables in the test set, as opposed to an unrepresentative relationship between variables in the whole of the data (overfitting), remains a problem. While the remedy used thus far involves the repeated testing of the same data, avoiding a single biased test set via repetitively creating many different ones, a dataset may be so large that an alternative approach requiring only a single run of the algorithm might be preferred.

There is a concern related to BeSiVa's operation that its single test set might prove unrepresentative of the whole of the data. Fortunately alternatives exist within the analytics literature that might prevent such unrepresentativeness. One example is known as k-fold cross-validation, where instead of designating a single test set, the observations are divided into what are known as folds (Kuhn and Johnson 2013). One fold of data is held out, while the others are used to estimate a model, similar to BeSiVa, but once the measures of model quality are estimated, the fold is replaced and another is removed. The model is estimated again, the held out fold is returned and another is removed, and this is done for all folds. The resulting measures of model quality (in the case of BeSiVa, the PCPs or PClPs generated for each fold) are averaged, and the model is selected based on the results of the average. While such an approach would need testing to determine how its results differed from BeSiVa's current operation, k-fold cross-validation would help prevent the overfitting that might trouble the algorithm in its current

incarnation.

Throughout the dissertation, the goal has been to demonstrate how BeSiVa may be used to develop results. One way to think of an algorithm like BeSiVa, however, is akin to an old car. While it is necessary to make sure that the algorithm can answer a research question, it is also important to perform additional maintenance and make improvements, to ensure that the algorithm runs well. Altering different aspects of the algorithm, in this case, would be akin to taking a car in for regular maintenance, fixing what isn't working well and maintaining what does work well. By changing elements of the algorithm's operation to make it run better, it ensures its continued use and expansion to other research questions. To wit, the car has taken us from place to place. It is now time to verify its continued operation, and make it work better for future trips.

### More than Oracles: The Fulfillment, Rather Than the Destruction Of Theory

The major contribution of this dissertation is an algorithm, and with it, the chance to review and create new findings in political science from a predictive standpoint. It is tempting to see the development of a predictive approach, however, as the end of theorizing within the social sciences, with predictive approaches designed around a problem related to human behavior, as an abolition of theory's role within the discipline. This might seem tenable, given that the algorithm regularly selected theoretically relevant predictors for the question of turnout, even when irrelevant predictors were included. Through predictive approaches and machine learning based techniques the discipline becomes capable of performing acts akin to that of an oracle, who sought to interpret the future through absurd techniques. This, however, runs exactly opposite

what was demonstrated by the results of this work. Despite providing irrelevant predictors to test the algorithm, BeSiVa's provided results always demanded a theoretical perspective for the purpose of understanding the findings.

What I hope this dissertation has demonstrated is that despite the need for additional methodological advancements to enhance the certainty surrounding the utility of a given finding, theory remains necessary to understand all of the results. In order to truly understand a phenomenon, it needs to be considered from a variety of angles, with theory either serving as the framework by which a concept of interest is understood, or the bounds by which the algorithm's predictors are selected, just like regression. For instance, predictors suggested a theory may prove to be statistically significant, but predictive testing using BeSiVa may reveal weakness in that theoretical connection. It may also be that a theory's predictive capability waxes and wanes depending on other contextual elements; maybe Rosenstone and Hansen's mobilization theory is better at predicting political participation closer to the time it was developed, as an example. No predictive algorithm may substitute for theoretical discernment in understanding what has changed between a theory's development and the present, or if it has changed at all. What the algorithm provides, however, is a chance to reconsider the relationship between theory and methods, making them each more equal in their ability to answer questions. It is through predictive techniques, and the BeSiVa algorithm specifically, that political science might leave the uncertainty of the purely explanatory approach behind for a lessened uncertainty provided by a combination of the explanatory and predictive approaches together. By proving theoretical relationships' substantive relevance through prediction, in addition to their statistical significance through explanation, we might go from assuming that theoretical relationships hold because of

the stars to proving that they do thanks to the predictions that they successfully make.

# Bibliography

## Introduction


Bartels, Larry M. 1997. "Specification Uncertainty and Model Averaging." *American Journal of Political Science* 41(April):641-674.

Clark, Todd E. 2004. "Can Out-of-Sample Forecast Comparisons Help Prevent Overfitting?" *Journal of Forecasting.* 23:115-139.

Cohen, Jacob. 1994. "The Earth is Round ($p < .05$)." *American Psychologist* 49:997-1003.

Cureton, Edward E. 1950. "Validity, Reliability, and Baloney." *Education and Psychological Measurement* 20:94-96.

Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(September):647-674.

Harrell, Frank. 1996. "What are some of the problems with stepwise regression?" May 1998. http://www.stata.com/support/faqs/statistics/stepwise-regression-problems/ (November 10, 2014)

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements Of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. New York: Springer.

Monroe, Burt L, Jennifer Pan, Margaret E. Roberts, Maya Sen, and Betsy Sinclair. 2014. "No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science." *PS* (January):71-74.

Mosteller Frederick and John W. Tukey. 1968. "Data Analysis, Including Statistics" in *The Handbook of Social Psychology Second Edition* Eds Gardner Lindzey and Elliot Aronson. Reading: Addison-Wesley Publishing.

Mosteller, Frederick and John W. Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics.* Reading: Addison-Wesley Publishing Company.

Pollard, P. and J.T.E. Richardson. 1987. "On the Probability of Making Type I Errors." *Psychological Bulletin* 102:159-163.

Schrodt, Philip. 2014. "Seven Deadly Sins of Contemporary Quantitative Political Analysis." *Journal of Peace Research.* 51:287-300.

Yu, Bei, Stefan Kaufmann, and Daniel Diermeier. 2008. "Classifying Party Affiliation from Political Speech." *Journal of Information Technology and Politics* 5:33-48.

## Chapter 1

Achen, Christopher H. 2002. "Towards a New Political Methodology: Microfoundations and ART." *Annual Review of Political Science.* 5:423-450.

Brody, Richard A. and Paul M. Sniderman.1977. "From Life Space to Polling Place: The Relevance of Personal Concerns for Voting Behavior." *British Journal of Political Science* 7(July):337-360.

Caldwell, Sally. 2012. *Statistics Unplugged* Belmont: Cengage Learning.

Campbell, Angus, Philip E. Converse, Warren E. Miller, and Donald E. Stokes. 1960. *The American Voter.* Chicago: University of Chicago Press, Midway.

Clark, Todd E. 2004. "Can Out-of-Sample Forecast Comparisons Help Prevent Overfitting?" *Journal of Forecasting.* 23:115-139.

Cohen, Jacob. 1994. "The Earth is Round ($p < .05$)." *American Psychologist* 49:997-1003.

Cureton, Edward E. 1950. "Validity, Reliability, and Baloney." *Education and Psychological Measurement* 20:94-96.

Gelman, Andrew and Eric Loken. 2013. "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time." Columbia University and Penn State University. Typeset.

Gerber, Alan, Donald P. Green, and Ron Shachar. 2003. "Voting May Be Habit-Forming: Evidence From a Randomized Field Experiment." *American Journal of Political Science* 47(July):540-550.

Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(September):647-674.

Grove, William M. and Paul E. Meehl. 1996. "Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical-Statistical Controversy." *Psychology, Public Policy, and Law* 2:293-323.

Haller, Heiko and Stefan Krauss. 2002. "Misinterpretation of Significance: A Problem Students Share with Their Teachers?" *Methods of Psychological Research Online* 7:1-20.

Harrell, Frank. 1996. "What are some of the problems with stepwise regression?" May 1998. http://www.stata.com/support/faqs/statistics/stepwise-regression-problems/ (November 10, 2014)

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements Of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. New York: Springer.

Hindman, Matthew. 2015. "Building Better Models: Prediction, Replication and Machine Learning in the Social Sciences." *The Annals of the American Association of Political And Social Science* 659(May): 48-62.

James Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R* New York: Springer.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry* Princeton: Princeton University Press.

Kuhn, Max and Kjell Johnson. 2013. *Applied Predictive Modeling.* New York: Springer.

Lewis-Beck, Michael. 2005. *Data Analysis: An Introduction* Thousand Oaks: Sage Publications.

Matloff, Norman. 2011. *The Art of R Programming.* San Francisco: No-Starch Press.

Meehl, Paul E. 1954. *Clinical Versus Statistical Prediction: A Theoretical Analysis and Review of Evidence.* Minneapolis: University of Minnesota Press.

Mosteller, Frederick and John W. Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics.* Reading: Addison-Wesley Publishing Company.

Rabe-Hesketh, Sophia and Anders Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata.* College Station: Stata Press.

Reinhart, Alex. 2015. *Statistics Done Wrong: The Woefully Complete Guide* San Francisco: No Starch Press.

Rogers, Benjamin J. 2014. "On the Creation of the BeSiVa Algorithm to Predict Voter Support" Master's thesis. The University of Kansas.

Schrodt, Philip. 2014. "Seven Deadly Sins of Contemporary Quantitative Political Analysis." *Journal of Peace Research.* 51:287-300.

Shapiro, Ian. 2002. "Problems, Methods, and Theories in the Study of Politics, or What's Wrong with Political Science and What to Do about It." *Political Theory.* 30(August) 596-619.

Shmuelli, Galit. 2010. "To Explain or to Predict?" *Statistical Science.* 25(August) 289-310.

Siegel, Eric. 2013. *Predictive Analytics: The Power to Predict Who will Click, Buy, Lie, or Die.* Hoboken: John Wiley & Sons.

Simmons, Joseph, Leif D. Nelson and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything As Significant." *Psychological Science* 22:1359-1366.

Teixeira, Ruy A. 1987. *Why Americans Don't Vote: Turnout Decline in the United States 1960-1984.* New York: Greenwood Press.

Verzani, John. 2005. *Using R for Introductory Statistics.* New York: Chapman & Hall/CRC.

Yom, Sean. 2015. "From Methodology to Practice: Inductive Iteration in Comparative Research." *Comparative Political Studies* 48:616-644.

## Chapter 2

Achen, Christopher H. 2002. "Towards a New Political Methodology: Microfoundations and ART." *Annual Review of Political Science.* 5:423-450.

Baumgartner, Frank R. and Beth L. Leech. 1998. *Basic Interests: The Importance of Groups in Politics and in Political Science* Princeton: Princeton University Press.

Belli, Robert, Michael W. Traugott, and Matthew N. Beckmann. 2001. "What Leads to Voting Overreports? Contrasts of Overreporters to Validated Voters and Admitted Nonvoters in the American National Election Studies." *Journal of Official Statistics.* 17:479-498.

Bendor, Jonathan, Daniel Diermeier, and Michael Ting. 2003. "A Behavioral Model Of Turnout." *The American Political Science Review* 97(May):261-280.

Berelson, Bernard, Paul F. Lazarsfeld, and William N. McPhee *Voting: A Study of Opinion Formation in a Presidential Campaign* Chicago: The University of Chicago Press.

Brody, Richard A. and Paul M. Sniderman.1977. "From Life Space to Polling Place: The Relevance of Personal Concerns for Voting Behavior." *British Journal of Political Science* 7(July):337-360.

Campbell, Angus, Philip E. Converse, Warren E. Miller, and Donald E. Stokes. 1960. *The American Voter.* Chicago: University of Chicago Press, Midway.

Clark, Todd E. 2004. "Can Out-of-Sample Forecast Comparisons Help Prevent Overfitting?" *Journal of Forecasting.* 23:115-139.

Cohen, Jacob. 1994. "The Earth is Round ($p < .05$)." *American Psychologist* 49:997-1003.

Dalton, Russell J. and Martin P. Wattenberg. 1993. "The Not So Simple Act of Voting." In *Political Science: The State of the Discipline II.* eds. Ada W. Finifter. Washington: The American Political Science Association.

Forster, Malcolm R. 2002. "Predictive Accuracy as an Achievable Goal of Science." *Philosophy of Science* 69(September):S124-S134.

Forster, Malcolm R. and Elliott Sober. 1994. "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions." *The British Journal for the Philosophy of Science* 45(March):1-35.

Fowler, James H. 2006. "Habitual Voting and Behavioral Turnout." *The Journal Of Politics* 68(May)335-344.

Gelman, Andrew and Eric Loken. 2013. "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time." Columbia University and Penn State University. Typeset.

Gerber, Alan, Donald P. Green, and Ron Shachar. 2003. "Voting May Be Habit-Forming: Evidence From a Randomized Field Experiment." *American Journal of Political Science* 47(July):540-550.

Hindman, Matthew. 2015. "Building Better Models: Prediction, Replication and Machine Learning in the Social Sciences." *The Annals of the American Association of Political And Social Science* 659(May): 48-62.

Kuhn, Max and Kjell Johnson. 2013. *Applied Predictive Modeling.* New York: Springer.

Long, J. Scott.1997. Regression *Models for Categorical and Limited Dependent Variables* Thousand Oaks: Sage Publications.

Mosteller, Frederick and John W. Tukey. 1977. *Data analysis and Regression: A Second Course in Statistics.* Reading: Addison-Wesley Publishing Company.

Olsen, Mancur. 1971. *The Logic of Collective Action* New Haven: Harvard University Press.

Rabe-Hesketh, Sophia and Anders Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata Volume II: Categorical Responses, Counts, and Survival.* College Station: Stata Press.

Reinhart, Alex. 2015. *Statistics Done Wrong: The Woefully Complete Guide* San Francisco: No Starch Press.

Rosenstone, Stephen J. and John M. Hansen. [1993] 2003. *Mobilization, Participation, and Democracy in America.* New York: Longman.

Schattschneider, E.E. 1975. *The Semisovereign People: A Realist's View of Democracy in America.* Hinsdale: Dryden Press

Schrodt, Philip. 2014. "Seven Deadly Sins of Contemporary Quantitative Political Analysis." *Journal of Peace Research.* 51:287-300.

Teixeira, Ruy A. 1987. *Why Americans Don't Vote: Turnout Decline in the United States 1960-1984.* New York: Greenwood Press.

Tourangeau, Roger and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin.* 5:859-883.

Verba, Sidney, Kay Lehman Schlozman, Henry Brady and Norman H. Nie. 1993. "Race, Ethnicity and Political Resources: Participation in the United States." *British Journal of Political Science.* 23(October):453-497.

Wolfinger, Raymond E. and Stephen J. Rosenstone. 1980. *Who Votes?* New Haven: Yale University Press.

## Chapter 3

Alba, Richard, Ruben G. Rumbaut, and Karen Marotz. 2005. "A Distorted Nation: Perceptions of Racial/Ethnic Group Sizes and Attitudes Toward Immigrants and Other Minorities." *Social Forces* 84(December):901-919.

American National Election Studies. 2016. "Codebook and User's Guide to the ANES 2016 Pilot Study." (April 24) http://www.electionstudies.org/studypages/anes_pilot_2016/anes_pilot_2016_CodebookUserGuide.pdf (February 23, 2016)

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees* Belmont: Wadsworth International Group.

Campbell, Angus, Philip E. Converse, Warren E. Miller, and Donald E. Stokes. 1960. *The American Voter.* Chicago: University of Chicago Press.

Carmines, Edward, Michael J. Ensley, and Michael W. Wagner. 2016. "Ideological Heterogeneity and the Rise of Donald Trump." *The Forum* 14:385-397.

Dougherty, Kevin D. 2003. “How Monochromatic Is Church Membership? Racial-Ethnic Diversity in Religious Community.” *Sociology of Religion* 64:65-85.

Enders, Adam M. and Steven Smallpage. 2016. “Racial Prejudice, Not Populism or Authoritarianism, Predicts Support for Trump over Clinton.” (May 26) https://www.washingtonpost.com/news/monkey-cage/wp/2016/05/26/these-9-simple-charts-show-how-donald-trumps-supporters-differ-from-hillary-clintons/ (March 2, 2017)

Fiorina, Morris P. 1981. *Retrospective Voting in American Elections* New Haven: Yale University Press.

Geisser, Seymour. 1974. “The Predictive Sample Reuse Method with Applications.” *Journal of the American Statistical Association* 70(June):320-328.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements Of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. New York: Springer.

James Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R* New York: Springer.

Kuhn, Max and Kjell Johnson. 2013. *Applied Predictive Modeling.* New York: Springer.

Kunovich, Robert M. 2016. “Perceptions of Racial Group Size in a Minority-Majority Area.” *Sociological Perspectives* 1-18.

Lawrence, Eric D. and John Sides. 2014. “The Consequences Of Political Innumeracy.” *Research and Politics* (July-September):1-8.

McLaren, Lauren M. 2003. “Anti-Immigrant Prejudice in Europe: Contact, Threat Perception, and Preferences for the Exclusion of Migrants.” *Social Forces* 81(March):909-936.

Monroe, Burt L, Jennifer Pan, Margaret E. Roberts, Maya Sen, and Betsy Sinclair. 2014. “No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science.” *PS* (January):71-74.

Rogers, Benjamin J. 2014. “On the Creation of the BeSiVa Algorithm to Predict Voter Support” Master's thesis. The University of Kansas.

Schaffner, Brian. 2016. “White support for Donald Trump was driven by economic anxiety, but also by racism and sexism.” March 28. http://www.vox.com/mischiefs-of-faction/2016/11/16/13651184/trump-support-economic-anxiety-racism-sexism (November 16, 2016)

Schrodt, Philip. 2014. “Seven Deadly Sins of Contemporary Quantitative Political Analysis.” *Journal of Peace Research.* 51:287-300.

Shapiro, Ian. 2002. “Problems, Methods, and Theories in the Study of Politics, or What's Wrong with Political Science and What to Do about It.” *Political Theory.* 30(August) 596-619.

Shmueli, Galit and Otto R. Koppius. 2011. “Predictive Analytics in Information Systems Research.” *MIS Quarterly* 35(September):553-572.

Stone, M. 1974. “Cross-validatory Choice and Assessment of Statistical Predictions.” *Journal of*

*the Royal Statistical Society. Series B(Methodological)* 36:111-147.

Yom, Sean. 2015. "From Methodology to Practice: Inductive Iteration in Comparative Research." *Comparative Political Studies* 48:616-644.

Conclusion

Achen, Christopher H. 2002. "Towards a New Political Methodology: Microfoundations and ART." *Annual Review of Political Science.* 5:423-450.

Alba, Richard, Ruben G. Rumbaut, and Karen Marotz. 2005. "A Distorted Nation: Perceptions of Racial/Ethnic Group Sizes and Attitudes Toward Immigrants and Other Minorities." *Social Forces* 84(December):901-919.

Brody, Richard A. and Paul M. Sniderman.1977. "From Life Space to Polling Place: The Relevance of Personal Concerns for Voting Behavior." *British Journal of Political Science* 7(July):337-360.

Clark, Todd E. 2004. "Can Out-of-Sample Forecast Comparisons Help Prevent Overfitting?" *Journal of Forecasting.* 23:115-139.

Cohen, Jacob. 1994. "The Earth is Round ($p < .05$)." *American Psychologist* 49:997-1003.

Cureton, Edward E. 1950. "Validity, Reliability, and Baloney." *Education and Psychological Measurement* 20:94-96.

Fiorina, Morris P. 1981. *Retrospective Voting in American Elections* New Haven: Yale University Press.

Fox, John. 2008. *Applied Regression Analysis and Generalized Linear Models* Los Angeles: Sage Publications.

Geisser, Seymour. 1974. "The Predictive Sample Reuse Method with Applications." *Journal of the American Statistical Association* 70(June):320-328.

Gelman, Andrew. 2008. "Objections to Bayesian Statistics." *Bayesian Analysis* 3:445–450.

Gelman, Andrew and Eric Loken. 2013. "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time." Columbia University and Penn State University. Typeset.

Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(September):647-674.

Harrell, Frank. 1996. "What are some of the problems with stepwise regression?" May 1998. http://www.stata.com/support/faqs/statistics/stepwise-regression-problems/ (November 10, 2014)

Kuhn, Max and Kjell Johnson. 2013. *Applied Predictive Modeling.* New York: Springer.

Meehl, Paul E. 1954. *Clinical Versus Statistical Prediction: A Theoretical Analysis and Review of Evidence.* Minneapolis: University of Minnesota Press.

Reinhart, Alex. 2015. *Statistics Done Wrong: The Woefully Complete Guide* San Francisco: No Starch Press.

Rosenstone, Stephen J. and John M. Hansen. [1993] 2003. *Mobilization, Participation, and Democracy in America.* New York: Longman.

Shmuelli, Galit. 2010. "To Explain or to Predict?" *Statistical Science.* 25(August) 289-310.

Silver, Nate. 2012. *The Signal and The Noise* New York:The Penguin Press.

Stone, M. 1974. "Cross-validatory Choice and Assessment of Statistical Predictions." *Journal of the Royal Statistical Society. Series B(Methodological)* 36:111-147.

Teixeira, Ruy A. 1987. *Why Americans Don't Vote: Turnout Decline in the United States 1960-1984.* New York: Greenwood Press.